tRNS-ek identitásvizsgálata új, in silico módszerrel

Doktori (PhD) értekezés

Szenes Áron

Témavezetők:
Dr. Pál Gábor docens és Dr. Jakó Éena tudományos főmunkatárs

Eötvös Loránd Tudományegyetem
Biológia Doktori Iskola
Vezetője: Prof. Dr. Erdei Anna, az MTA levelező tagja

Szerkezeti Biokémia Doktori Program
Vezetője: Prof. Dr. Gráf László, az MTA rendes tagja

2012.

Így szóltam magamban: „Szeretnék szert tenni a bölcsességre!" De távol maradt tőlem. Ami van, messze van, és mélységes mélyen! Ki tudja megtalálni?[...]
Amikor azon fáradoztam, hogy megismerjem a bölcsességet, [...]
akkor láttam: minden az Isten műve, s az ember nem képes felfogni az eseményeket, amelyek a nap alatt lejátszódnak. Bármennyit fárad is az ember a kutatással, nem jut el a megértésig. És maga a bölcs sem tud a dolog nyitjára jönni, aki azt gondolja, hogy érti.
(Préd. 7,23-24; 8,16-17)

Köszönetnyilvánítás

Köszönettel tartozom témavezetőimnek, Jakó Éenának, aki megalkotta az ECP módszert, illetve Pál Gábornak, akinek kritikáira és angoltudására mindig számíthattam.

Köszönet illeti a munkában résztvevő társszerzőket, Ittzés Pétert, aki a statisztikai szoftverek egy részét készítette, Kun Ádámot, aki szintén statisztikai elemzésekben vett részt, Szathmáry Eörsöt, aki ötleteivel segítette és rendszerezte a munkát valamint Horváth Arnoldot, aki az ECP-t futtató egyik programot készítette.

Köszönöm Gráf Lászlónak, hogy iskolateremtő munkája megfelelő szakmai hátteret biztosított a dolgozat megszületéséhez, illetve Nyitray Lászlónak és Juhász Gábornak, hogy dolgozhattam csoportjukban.

Köszönöm Catherine Florentznek a munkához adott általános segítségét és észrevételeit, Szenes Márknak pedig a statisztikai módszerekben és matematikai formalizmusokban nyújtott segítségét.

Kiemelt köszönet illeti meg Barta Endrét, aki megtanított a bioinformatikai módszerek alapjaira. Nélküle ez a munka nem jöhetett volna létre.

Összefoglaló

Minden élő szervezetben kiemelten fontos, hogy az aminoacil-tRNS szintetázok (aaRS) a meg felelő tRNS molekulát ismerjék föl, és a DNS-en tárolt információ a genetikai kódnak megfelelő módon a fehérjeszintézis során hiba nélkül, pontosan fejeződjön ki. Az aaRS enzimeket két osztályba sorolhatjuk szekvenciájuk és térszerkezetük különbözősége alapján. Ezen különbségeknek, amelyek az élővilág mindhárom nagy csoportjában meggyelhetőek, feltehetően evolúciós okai vannak. A szintetázok felosztását követve, annak analógiájára, az általuk aminosavval feltöltött tRNS-eket is besorolhatjuk osztályokba. A tRNS-szintetáz kapcsolat a legtöbb esetben erősen specikus, ezért feltételezhetjük, hogy egyfajta koevolúciós folyamat során a tRNS szekvenciában is megmaradt a nyoma annak, hogy melyik osztályba tartozik a szintetáza. Az irodalomban eddig ismert adatok szerint azonban ilyen osztályspecikus szekvenciaelemek nem léteznek. Jakó Éena ennek megvizsgálása érdekében megalkotott egy új, diszkrét matematikai módszert, amely az egyes pozíciókat vizsgálva, nem csak a minden szekvenciában jelen levő, hanem az összes szekvenciából hiányzó nukleotidokat is gyelembe veszi. Az ECP analízis eredményéül az ún. „diszkrimináló elem"-eket (DE) kapjuk. A DE az a bázis, vagy azok a bázisok, amelyek az egyik osztály adott pozíciójában minden szekvenciából hiányoznak, de a másik osztály ugyanazon pozíciójában a szekvenciák közül legalább egyben megtalálhatóak. Munkánk során 50 faj (13 ősbaktérium, 30 baktérium és 7 eukarióta) I. és II. osztályú tDNS szekvenciáján elvégezve az ECP analízist, azt találtuk, hogy léteznek osztályspecikus DE-k, amelyeket az eddig alkalmazott módszerekkel nem sikerült feltárni. Statisztikai módszerekkel igazoltuk, hogy az osztályok szekvencia-alapú szétválasztására az ECP hatékonyabb az eddigi megközelítéseknél. Az osztályspecikus, bakteriális tDNS szekvenciákra jellemző DE-ket reprodukáltuk egy nem diszkrét, Shannon-entrópián alapuló módszerrel is. Az ECP módszert továbbfejlesztve egy újabb eljárást terveztünk annak érdekében, hogy a tDNS szekvenciák között a 20 -féle identitás között tegyünk különbséget. Ehhez szűrt adatbázisokat készítettünk, amelyhez a szűrési szempontokat az ismert, minden egyes tRNS molekulára jellemző tulajdonságok, illetve a már publikált identitáselemek jelentették. Az analízist az élővilág mindhárom nagy csoportjában elvégeztük. Az ECP analízis DE-it kiszámítottuk minden pozícióban, minden egyes aminosavidentitású tDNS csoportot mindegyik mással párba állítva, összesen 380 párt képezve, majd minden pozícióban megállapítottuk a DE-k átlagos számát. Ezt az értéket neveztük el átlagos kizárási értéknek („*average excluding value*", AEV). Az AEV értékeit pozíciónként összehasonlítottuk a már publikált identitáselemekkel, és statisztikai módszerekkel igazoltuk, hogy – az adatbázis szűrésétől függetlenül – a két érték korrelál egymással, azaz a magas AEV értékű pozíciók feltehetően identitáselemeket hordoznak.

Abstract

In all organisms, the 20 aminoacyl-tRNA synthetase (aaRS) enzymes have to recognize their amino acid substrates and the corresponding tRNA molecules with high precision to produce only legitimate aminoacyl-tRNA products. This exquisite specificity is of central importance as this enables the genetic information to be faithfully translated into protein sequences by following the rules defined in the genetic code. aaRS are grouped into Class I and II based on primary and tertiary structure and enzyme properties suggesting two independent phylogenetic lineages. Analogously, tRNA molecules can also form two respective classes, based on the class membership of their corresponding aaRS. Although some aaRS–tRNA interactions are not extremely specific and require editing mechanisms to avoid misaminoacylation, most aaRS–tRNA interactions are rather stereospecific. Thus, class-specific aaRS features could be mirrored by class-specific tRNA features. However, previous investigations failed to detect conserved class-specific nucleotides. Éena Jakó introduced a discrete mathematical approach that evaluates not only class-specific 'strictly present', but also 'strictly absent' nucleotides. The disjoint subsets of these elements compose a unique partition, named extended consensus partition (ECP). The ECP identifies nucleotide types at each position that are strictly absent from a given sequence set, while occur in other sets. These are defined as discriminating elements (DEs). By analyzing the ECP for both Class I and II tDNA sets from 50 (13 archaeal, 30 bacterial and 7 eukaryotic) species, we could demonstrate that class-specific DEs do exist, although not in terms of strictly conserved nucleotides as it had previously been anticipated. This finding demonstrates that important information was hidden in tRNA sequences inaccessible for traditional statistical methods. With an information-theory based, non-discrete method, we reproduced the results of ECP analysis in bacterial dataset. Using the ECP approach, we mapped potential hidden identity elements that discriminate the 20 different tRNA identities. We filtered the tDNA data set for the obligatory presence of well-established tRNA features, and then separately for each identity set, the presence of already experimentally identified strictly present identity elements. The analysis was performed on the three kingdoms of life. We determined the number of DE, e.g. the number of sets discriminated by the given position, for each tRNA position of each tRNA identity set. Then, from the positional DE numbers obtained from the 380 pairwise comparisons of the 20 identity sets, we calculated the average excluding value (AEV) for each tRNA position. The AEV provides a measure on the overall discriminating power of each position. Using a statistical analysis, we show that positional AEVs correlate with the number of already identified identity elements. Positions having high AEV but lacking published identity elements predict hitherto undiscovered tRNA identity elements.

Tartalomjegyzék

1. Bevezetés 1

1.1. A tRNS-ek biológiai szerepe 1

1.2. A tRNS identitás fogalma és biokémiai háttere 2

1.2.1. Az aminoacil-tRNS szintetázok osztályai 2

1.2.2. A tRNS-ek és az aminoacil-tRNS szintetázok kapcsolata – az identitáselemek típusai 3

1.2.3. Az identitáselemek elhelyezkedése a tRNS molekulán 3

1.2.4. Az antideterminánsok 6

1.3. A tRNS-ek identitásvizsgálatának perspektívái 6

1.4. Az identitáselemek kísérletes meghatározása 7

1.5. Az in silico identitásvizsgálat lehetőségei 10

1.5.1. tRNS-ek in silico meghatározása genomi szekvenciákon 10

1.5.2. tRNS adatbázisok 11

1.5.3. A nukleotidok és azok csoportjainak IUPAC jelöléskonvenciója 12

1.5.4. A szekvencia „*logo*”-k 13

1.6. A makromolekula-funkciók feltárásának általános elvei 15

2. Célkitűzések 18

3. Módszerek 19

3.1. Programok és programnyelvek 19

3.2. Felhasznált adatbázisok jellemzői 20

3.2.1. A tRNomics adatbázis 20

3.2.2. Az MSDB adatbázis 20

3.2.3. A tDNAdbC adatbázis 21

3.3. Az ECP algoritmus 21

3.3.1. Az SCP algoritmus 21

3.3.2. Az ECP rövid, mesterséges szekvenciákon 22
3.4. Statisztikai módszerek 24
3.4.1. Az ECP hatékonyságának tesztelése 24

4. Módszerfejlesztés **26**
4.1. Az adatbázisok átalakítása; saját, szűrt adatbázisok készítése 26
4.1.1. A tRNomics feldolgozása 26
4.1.2. Az MSDB feldolgozása 27
4.1.3. A tDNAdbC szűrése 27
4.2. Az ECP használata tRNS-identitásokra 29
4.2.1. Az AEV 30
4.2.2. Az ECP módszer és az AEV formalizálása 30

5. Eredmények és értelmezésük **33**
5.1. A tRNS szekvenciák szekvencia alapú szétválasztása szintetáz osztályuknak megfelelően ECP módszerrel 33
5.1.1. Az ECP tRNS/tDNS szekvenciákon 33
5.1.2. Az SCP és ECP összehasonlítása 35
5.1.3. Az ECP analízis osztályspecifikus diszkrimináló elemei 38
5.1.4. Az ECP osztályokat szétválasztó képessége 41
5.1.5. Egyedi, osztályspecifikus DE-készletek 43
5.1.6. Egyedi DE-k 43
5.1.7. Az ECP módszer értékelése 45
5.1.8. Az osztályspecifikus elemek kísérleti eredmények tükrében 45
5.2. Osztályspecifikus elemek feltárása *„logo”* módszerrel 47
5.2.1. Az I. és a II. osztály *„inverse function logo”*-i 47
5.2.2. Az *„inverse function logo”*-k és a diszkrimináló elemek összefüggései . . 48
5.2.3. Az I. és a II. osztály *„logo”*-inak értékelése 48
5.3. Új identitás helyek feltérképezése tRNS pozíciók átlagos DE számának segítségével 50
5.3.1. Az AEV statisztikai értékelése 50
5.3.2. Az AEV eredményei 53
5.3.3. Eukarióta (élesztő) adatok 56
5.3.4. Az adatszűrés lehetséges hatása az eredményekre 57
5.3.5. Ősbakteriális adatok 59
5.3.6. Potenciális identitáselemek 62

5.4. Konklúzió 67

Rövidítések jegyzéke

Aac-RS: adott aminosav identitású transzfer ribonukleinsav szintetáz enzime, ahol Aac az aminosav hárombetűs kódja

aaRS: aminoacil-transzfer ribonukleinsav szintetáz enzim

AEV: „*average excluding value*”, átlagos kizárási érték

CAEV: „*cumulative average excluding value*”, egy pozícióhoz tartozó átlagos kizárási értékek összege

DE: „*discriminating elements*”, diszkrimináló elemek

ECP: „*extended consensus partition*”, egyedi osztályozó módszer, amely a „*strictly absent*” elemeket is figyelembe veszi

GtRNAdb: ‘„*Genomic tRNA Database*”

NPD: „*number of published determinants*”, egy pozícióban található, az irodalomban már publikált identitáselemek összege

SA: „*strictly absent*” elemek, azok a nukleotidok, amelyek minden szekvenciában hiányoznak az adott pozícióban

SCP: „*strict consensus partition*”, osztályozó módszer, amely csak a „*strictly present*” elemeket veszi figyelembe

SP: „*strictly present*” elemek, azok a nukleotidok, amelyek minden szekvenciában jelen vannak az adott pozícióban

tDNS: transzfer ribonukleinsav géne

tRNAdb: „*tRNA database*”, transzfer ribonukleinsav adatbázis

tRNAdb-CE: „*tRNA Gene DataBase Curated by Experts*”, kézzel ellenőrzött transzfer ribonukleinsav adatbázis

tRNAdbC: saját fejlesztésű transzfer ribonukleinsav adatbázis

tRNS: transzfer ribonukleinsav

$tRNS^{Aac}$: adott aminosav identitású transzfer ribonukleinsav, ahol Aac az aminosav hárombetűs kódja

Táblázatok jegyzéke

1.1. Az I. és a II. aaRS osztálynak megfelelő tRNS-ek identitáselemei 5
1.2. Az antideterminánsok (Giegé nyomán, módosítva) 6
1.3. Az IUPAC jelöléskonvenciója (Sebestyén Endre nyomán) 12

5.1. A tDNS osztalyozás hatekonyságának matematikai analízise 37
5.2. Az osztályok jellemző SA („*strictly absent*”) elemei 44
5.3. A különböző adathalmazok mérete, illetve az elvégzett statisztikai analízisek eredményei 58
5.4. Kísérletesen megállapított ősbakteriális identitáselemek 61

Ábrák jegyzéke

1.1. tRNSArg és Arg-RS komlexe, PDB: 1F7V [1] 4
1.2. Az identitáselem-meghatározás *in vivo* és *in vitro* módszereinek összehasonlítása. 8
1.3. Az identitáselem-meghatározás *in vivo* sémájának részletes bemutatása. 9
1.4. tRNS-ek *„function logo"*-ja [2] 14

3.1. Az ECP működése rövid, mesterséges szekvenciákon 23

4.1. Az átlagos kizárási érték számítása rövid, mesterséges szekvenciákon. 31

5.1. Az ECP algoritmus működése az élesztő tDNS szekvenciáin 33
5.2. Élesztőből származó adatokkal végzett ECP analízis eredménye a tRNS két dimenziós szerkezetén 36
5.3. Az ECP analízis diszkrimináló elemei 38
5.4. Az ECP analízis diszkrimináló elemei az élővilág három nagy doménje szerint bontva 42
5.5. Az I. és a II. osztály bakteriális szekvenciáinak *„inverse function logo"*-ja 49
5.6. Az AEV értékek korrelációja az ismert identitáselemek számával 51
5.7. Az AEV értékek korrelációja az ismert identitáselemek számával 52
5.8. Az AEV és NPD értékei az élővilág három nagy doménjében 53
5.9. A bakteriális (A és B) és az eukarióta (C és D) adatok eredményei a második szűrési lépés kihagyásával 60
5.10. tRNSAsp – AspRS komplex szerkezetek 65
5.11. Lehetséges, eddig nem ismert identitáselemek 66

1.

Bevezetés

1.1. A tRNS-ek biológiai szerepe

A genetikai információ áramlásának, a fehérjeszintézisnek egyik kulcsszereplője a tRNS molekula. A fehérjeszintézis „szerelőasztalán", a riboszómán az mRNS kodonjával párba álló, megfelelő antikodonú aminoacil-tRNS molekula gondoskodik arról, hogy a DNS-ben tárolt, a genetikai kódnak megfelelő aminosav épüljön be a készülő fehérjébe. A tRNS-ek aminosavval való feltöltéséért az aminoacil-tRNS szintetáz enzimek (aaRS) felelősek, amelyek a tRNS-ek térszerkezetén megfelelő pozíciókat fölismerve a tRNS-ekhez azok antikodonjának megfelelő aminosavat kapcsolnak. Ez a felismerés, a helyes kapcsolódás tehát alapvető jelentőségű a helyes genetikai információ érvényre juttatásában az élővilág minden ágában [3, 4]. A felismerés, a helyes tRNS-aaRS szintetáz kapcsolat kulcsfogalma az identitás. Egy-egy tRNS identitása nem más, mint a szintetáz által hozzá kapcsolt, a tRNS antikodonjának megfelelő aminosav. A tRNS-nek azon nukleotidjait, amelyek fontosak abban, hogy kizárólag a megfelelő aaRS enzim ismerje fel az adott tRNS-t, a tRNS identitáselemeinek nevezik. Noha logikai alapon azt hihetnénk, hogy az aaRS enzimek kizárólag a tRNS antikodonját ismerik fel, a helyzet ennél jóval összetettebb. Bár a genetikai kód alapján az antikodon szekvenciájából egyértelműen meg tudjuk adni minden természetben előforduló tRNS identitását, az enzimek nem hagyatkozhatnak pusztán erre. Egyrészt a hasonló antikodonok megkülönböztethetősége nem lenne elegendő, másrészt vannak olyan aminosavak, amelyek 6-féle kodonnal bírnak, és az ezeket kiolvasó antikodonok egymástól nagymértékben eltérnek. Ebből következően, bár a tRNS-ek javarészében az enzim az antikodon bázisait is felismeri identitáselemként, az identitáselemek az antikodontól távolesó részeken is lehetnek, és vannak olyan tRNS-ek, ahol az antikodon bázisai egyáltalán nem szolgálnak az aaRS számára felismerőhelyként.

A genetikai kód univerzális jellege mellett a fehérjeszintézis alapfolyamata az élővilág minden egyes fajában azonos. Mindemellett a tRNS-szintetáz kapcsolat fajonként eltérő. Ez azt jelenti, hogy egy adott faj például alanin identitású tRNS-ét nem fogja bármelyik tetszőlegesen kiválasztott más faj alanin aaRS enzime felismerni. A tRNS-enzim kapcsolat szereplői, az egyes nukleotid bázisok a tRNS-en, tehát a tRNS identitáselemei és a szintetáz enzim aminosav csoportjai, azok egymással kialakított kapcsolódásai tehát fajonként is eltérhetnek. Ráadásul az eltérő identitású tRNS-ek sem feltétlenül azonos régiókban hordoznak (természetesen egymástól eltérő) identitáselemeket, maguk a régiók is eltérhetnek egymástól. Végül még az is megeshet, hogy egy adott fajon belül is olyan tRNS-ek, amelyek azonos identitásúak, egymástól részben eltérő helyeken hordozhatják az identitás elemeiket. Az eltérések okai, következményei, hatása a fehérjeszintézis menetére kiemelt jelentőségű kérdések és sok tekintetben a mai napig megválaszolatlanok.

1.2. A tRNS identitás fogalma és biokémiai háttere

1.2.1. Az aminoacil-tRNS szintetázok osztályai

Az aminoacil-tRNS szintetázokat alapvetően két csoportra tudjuk osztani szekvenciamintázataik, aktív centrumuk térbeli struktúrája valamint a aminosavkötő helyük különbözősége alapján [5–10]. A két osztály az I. és II. aaRS osztály, amely az élővilág minden csoportjában megtalálható [11–14]. Ezen különbségeknek evolúciós okai vannak: a két enzimcsalád egy-egy ős szintetáz enzimből fejlődhetett ki, a családok (a két osztály) pedig – megfelelő identitások esetén – ugyanúgy megfigyelhetőek az élővilág három nagy kingdomjában (az eukarióta, bakteriális és ősbakteriális csoportokban). A két osztályba mindhárom csoportba alapvetően ugyanaz a tíz-tíz aminosav-identitás tartozik, egyetlen kivétellel: a lizin-specifikus aminoacil-tRNS szintetáz (LysRS) mindkét osztályban előfordulhat [15–18] (bár minden konkrét faj vagy csak az egyik, vagy csak a másik osztályba tartozót hordozza). Ez a gyakorlatban azt jelenti, hogy a különböző osztályba tartozó LysRS enzimek ugyanúgy funkcionálnak (a megfelelő tRNSLys-t lizinnel töltik föl), de szekvenciájuk és ez által kialakított térszerkezetük eltér, a két külön osztály tulajdonságainak megfelelően [19, 20].

A szintetázok felosztását követve, annak analógiájára, az általuk aminosavval feltöltött tRNS-eket is besorolhatjuk osztályokba. A továbbiakban az „I. és II. osztály" megjelöléseket így a tRNS (vagy az őket kódoló gének, a tDNS-ek) szekvenciáira is használom. Itt megjegyzendő azonban, hogy ez az elnevezés nem összekeveredő az irodalomban használatos *„type I"* és *„type II"* tRNS típusok megjelölésével, amely a variábilis régió hosszára vonatkozó felosztást jelent

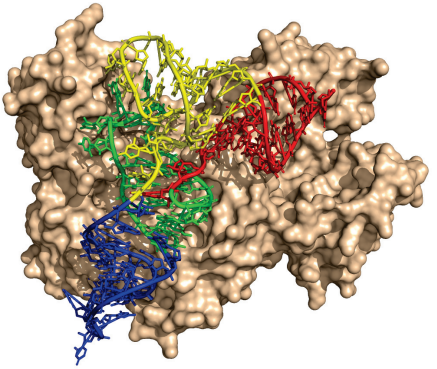
[21].
Jelen munka kezdetekor az irodalomban általánosan elfogadott feltételezés szerint az I. illetve II. osztályba tartozó tRNS-eket illetve a kódoló tDNS géneket szekvenciális alapon nem lehetett egymástól elválasztani, azaz nem léteztek osztályspecifikus szekvencia tulajdonságok, bázisok [22]. Az azonban ismert volt, hogy amennyiben a tRNS osztályokat tovább osztjuk aminosav-specifitásuk szerint, akkor az egyes izoakceptor-csoportokban már megjelennek az egyes identitásokra jellemző szekvenciális hasonlóságok [23–25].

1.2.2. A tRNS-ek és az aminoacil-tRNS szintetázok kapcsolata – az identitáselemek típusai

Az irodalmi áttekintés és munkám egésze során nagyban támaszkodtam Richard Giegé és strasbourgi csoportjának eredményeire. Kiváló összefoglaló művükben[24] megállapították a tRNS-ek identitására vonatkozó legfontosabb általános szabályokat és egyedileg jellemző tulajdonságokat. Kutatásaim során egyfajta zsinórmértéket jelentettek az itt leírtak, amelyek – noha a szerzők szerint is újabb kiegészítésekre szorulnak –, ma is megállják a helyüket. A szintetáz-tRNS kapcsolat, illetve az identitáskutatás egyidős a tRNS-ek felfedezésével [26–29]. E kapcsolatnak vannak úgynevezett „pozitív" illetve „negatív" elemei, a determinánsok illetve az antideterminánsok. Előbbiek funkciója az, hogy a tRNS-t a számára megfelelő aaRS ismerje fel, utóbbiak pedig az, hogy „távol tartsák maguktól" a számukra nem megfelelő szintetázokat, megakadályozva, hogy az illető tRNS tévesen, hibás aminosavval töltődjön föl. Az irodalom sokszor megkülönbözteti az tRNS identitásra szempontjából kiemelten fontos („*major*") és kisebb szerepet játszó („*minor*") elemeit . Előbbiek az identitást egyértelműen befolyásolják, létük vagy nem létük, (mutációjuk) hibás aminosav-feltöltést okoz. A kisebb elemek a finomhangolásért felelnek, hatásuk kisebb, általában csak a feltöltődés kinetikájára hatnak. Szintén különbséget tesznek az *in vivo* munkák során megállapított identitáselemek és az *in vitro* kísérletekből származó identitáselemek között (ez utóbbiak neve „*recognition elements*"), Giegé nyomán viszont jelen dolgozatban én sem használom ezt a különbségtételt.

1.2.3. Az identitáselemek elhelyezkedése a tRNS molekulán

A legegyszerűbb és legkézenfekvőbb megállapításokat a tRNS-aaRS kapcsolat feltárására akkor tehetjük meg, ha a komplex térszerkezete ismert, és azt tanulmányozzuk. A szerkezet meghatározása azonban közel sem egyszerű. Az egyik ismert szerkezetet a 1.1 ábrán mutatom be. Ahogyan már az előző fejezetben megemlítettem, identitáselemek találhatóak az antikodon



1.1. ábra. tRNSArg és Arg-RS komlexe, PDB: 1F7V [1]
Világosbarna, térkitöltő modellel ábrázoltam a szintetázt, szalag- és pálcikamodellel a tRNS molekulát, amelynek az egyes, két dimenziós (ún. „lóhere") szerkezetében jellegzetes régióit különböző színekkel kiemeltem: az akceptor-kar piros, az antikodon-hurok sárga, a D-hurok zöld a T-hurok sárgával jelölt.

hurkon: szinte az összes identitásnál a felismerésben fontos szerepet játszanak az antikodon bázisai. A másik fontos régió a tRNS „nyaki" része az ún. *„acceptor"* kar, és annak 3' végén, a -CCA (aminosavkötő) szekvencia előtt közvetlenül található diszkriminátor bázis. E két legjellemzőbb helyen kívül a molekula más részein is találhatunk identitáselemet. Az tRNS-ek identitásvizsgálatának alanyául alig pár modell-szervezet szolgál. Az *Escherichia coli* mellett az élesztő (*Saccharomyces cerevisiae*) valamint a *Thermus thermohilus* baktérium és néhány eukarióta (köztük az emberi) valamint ősbakteriális rendszert tanulmányoztak kísérletes módszerekkel. Mint említettem, az identitáselemek elhelyezkedése egy-egy aminosavidentitás esetén a különböző fajoknál akár el is térhet. Természetesen a különböző fajok (és az élőlények különböző csoportjaira vonatkozó általános szabályok) esetén egy identitáselem azonos pozícióban is gyakran más és más lehet. Fontos megjegyezni, hogy a módosított bázisok is a fent említett eltéréseket mutatják. Az identitás meghatározásában az *E. coli* baktériumban az izoleucin, glutaminsav és a lizin esetében az élesztőnél pedig az izoleucin esetében játszanak szerepet ilyen bázisok.

A felsorolt modellfajok tekintetében Giegé és munkatársai adták a legátfogóbb képet a feltárt identitáselemekről, amelyet az *E. coli* és az élesztő esetében az 1.1. táblázatban mutatok be. Itt jegyzendő meg, hogy a tRNS pozíciók számozása 0-től 73-ig, 5' $\rightarrow$ 3' irányban történik. A

1.1. táblázat. Az I. és a II. aaRS osztálynak megfelelő tRNS-ek identitáselemei

	I. osztály			II. osztály	
	E. coli	*S. cerevisiae*		*E. coli*	*S. cerevisiae*
Val	A73 G3:C70, T4:A69 A35, C36	A73 A35	**Ser**	G73 C72, G2:C71, A3:T70, C11:G24, R4:Y69	
Ile	A73 C4:G69 G34, A35, T36 A37, A38 T12:A23, C29:G41	G34, A35, T36	**Thr**	G1:C72, C2:G71 G34, G35, T36	G1:C72 G35:T36
Leu	A73 T8·A14	A73 A35 G37	**Pro**	A73 G72 G35, G36 G15C48	
Met	A73 T4:A69, A5:T68 C34, A35, T36	A73 C34, A35, T36	**Gly**	T73 G1:C72, C2:G71, G3:C70 C35, C36	A73 C2:G71, G3:C70 C35, C36
Cys	T73 G2:C71, C3:G70 G34, C35, A36	T73	**His**	C73 G0	A73 G0 G34, T35
Tyr	A73 T35	A73 C1:G72 G34, T35	**Asp**	G73 G2:C71 G34, T35, C36 C38 G10	G73 G34, T35, C36 C38 G10·T25
Trp	G73 A1:T72, G2:71 G3:C70 C34, C35, A36	C34, C35	**Lys**	A73 T34, T35, T36	
Glu	G1:C72, T2:A71 T34, T35 A37 T11:A24, T13:G22·A46, Δ47		**Asn**	G73 C34, T35, T36	
Gln	G73 T1:A72, G2:C71 G3:C70 Y34, T35, G36 A37, T38 G10		**Phe**	A73 G34, A35, A36 G27:C43, G28:C42 T20 G44, T45, T59, T60	A73 G34, A35, A36 A37 G20
Arg	A/G73 C35, T/G36 A20	C35, T/G36	**Ala**	A73 G2:C71, G3·T70 G4:C69 G20	G3·T70

1.2. táblázat. Az antideterminánsok (Giegé nyomán, módosítva)

Antideterminán	Melyik tRNS-en	Melyik aaRS-el szemben	Hivatkozás
G1	szintetikus $tRNA^{Gln}$ (*E. coli*)/I	TrpRS/I	[30]
G2•U71	$tRNA^{Lys}$ (*B. burgdorferi*)/II	LysRS/I (*E. coli*)	[31]
C4•G69	$tRNA^{Ile}$ (*E. coli*)/I	MetRS/I	[32],[33]
C31•G39	$tRNA^{Gln}$ (*E. coli*)/I	LysRS/I	[34]
U31•A39			
A5•U68	Ser-$tRNA^{Ser}$	PSTK (Ser-$tRNA^{Sec}$ kináz	[35]
	(*Methanococcus maripaludis*)/I	*Methanocaldococcus jannaschii*)	
G3•U70	$tRNA^{Ala}$ (élesztő)/II	ThrRS/II	[36]
U30•G40	$tRNA^{Ile}$ (élesztő)/I	GlnRS/I	[37]
		LysRS/I	
U34	$tRNA^{Ile}$ (élesztő)/I	MetRS/I	[38]
L34	$tRNA^{Ile}$ (*E. coli*)/I	MetRS/I	[39]
A36	$tRNA^{Trp}$ (*E. coli*)/I	ArgRS/I	[40]
A73	$tRNA^{Leu}$ (ember)/I	SerRS/II	[41]
G35	$tRNA^{Ser}$ (ember)/II	LeuRS/I	[41]
m^1G37	$tRNA^{Asp}$ (élesztő)/II	ArgRS/I	[42–44]
G73	$tRNA^{Ser}$ (élesztő)/II	LeuRS/I	[45]

variábilis hurok számozását „e" betű különbözteti meg, a már nevesített diszkriminátor bázis a 73-as. A már bemutatott ábrákon az egyes karok egységes színezéssel jelennek meg, a tRNS valódi, L alakú térszerkezete pedig a 1.1 ábrán is látható.

1.2.4. Az antideterminánsok

Ahogyan a determinánsok, úgy az antideterminánsok vizsgálata is szerepet kapott a kísérletes kutatások során, bár ez utóbbiakról jóval kevesebb eredményt publikáltak. A 1.2. táblázatot, amelyet (irodalmi kutatómunkám alapján, amely módosításokat később maga a szerző is elfogadott) kissé módosítottam, Giegé nyomán közlöm [24].

1.3. A tRNS-ek identitásvizsgálatának perspektívái

Miért lehet érdekes az identitásvizsgálat, milyen eredményeket hozhat a tRNS-ek identitáselemeinek feltárása? Azon túl, hogy egy ilyen alapvető lépés megismerése a fehérjeszintézisben önmagában érdekes, potenciálisan gyakorlati jelentőségű is lehet. A széleskörűen használt anti-

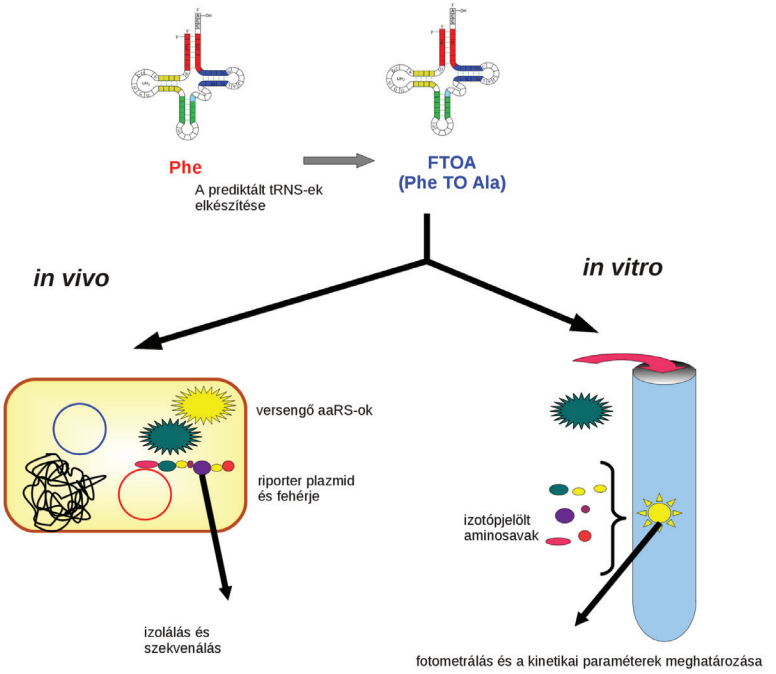
biotikumok hatékonysága a kórokozók rezisztenciájának kialakulása miatt rohamosan csökken (Schimmel 1998). Ezért az új típusú antibiotikumok kifejlesztése, amelyek új célpontok ellen hatnak, világszerte kiemelten fontos feladat. Ismert, rendkívül ígéretes, de még ki nem aknázott antibiotikum célpontok a mikrobák aminosav-tRNS szintetáz enzimei. Rendkívül fontos követelmény, hogy az antibiotikum által támadott mikrobiális szintetáz nagymértékben eltérjen az emberi szervezetben jelenlévő megfelelőjétől, annak érdekében, hogy az antibiotikum ne gátolja az emberi enzimet. Ha ez nem teljesül, akkor az antibiotikum nem használható, hiszen alkalmazása súlyos mellékhatásokkal járna. Célul tűzhető ki, hogy a tRNS-ekben kimutassuk a szintetáz felismerésben legfontosabb szekvencia-elemeket (identitáselemeket). A mikroba illetve emberi tRNS-készletek összehasonlító analízisével azonosíthatók azon tRNS-ek, amelyek leginkább eltérnek egymástól a két fajban. Az ezekhez tartozó szintetázok lehetnek a legmegfelelőbb antibiotikum célpontok.

1.4. Az identitáselemek kísérletes meghatározása

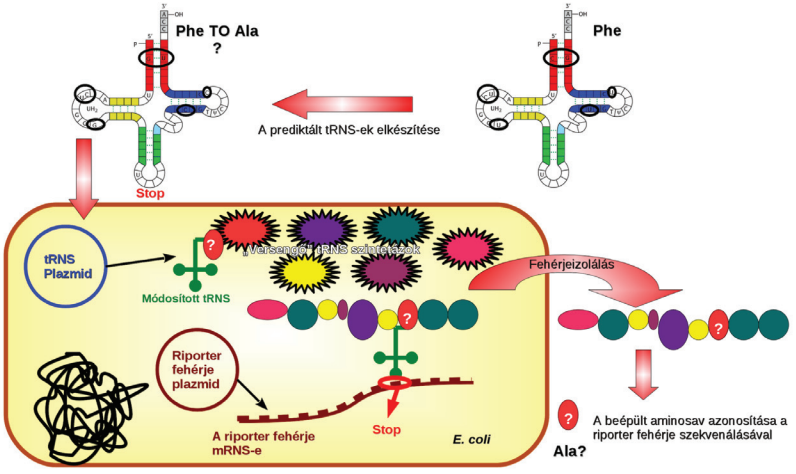
Az identitáselemek kísérletes meghatározása során azt igyekeznek felderíteni, hogy melyek azok a nukleotid pozíciók, amelyeknek – lehetőleg minimális számú – megváltoztatása megváltoztatja a tRNS identitását. Amikor ilyen kísérleteken keresztül identitáselemnek tűnik egy-egy nukleotid, akkor az identitás elem funkció legnyilvánvalóbb bizonyítéka az, ha az adott elemet egy másik tRNS-be áttéve áttevődik az identitás is. A megváltoztatott szekvenciájú tRNS identitása alapvetően kétféle kísérletes módszerrel vizsgálható meg. Az egyik egy *in vitro*, a másik egy *in vivo* rendszer. (Összehasonlítását lásd az 1.2 ábrán is.)

A két megközelítés kiegészíti egymást. Az *in vitro* rendszerben izolált tRNS és izolált aaRS enzim, valamint ATP és izotópjelölt aminosav felhasználásával játszatjuk le az aminoacilálás reakcióját. A keletkező aminoacil-tRNS koncentrációjának időbeni változását követve meghatározzuk az enzimreakcióra jellemző kinetikai paramétereket. A módszer előnye, hogy segítségével tetszés szerint bármilyen tRNS-enzim kombináció vizsgálható, és részletes kvantitatív paramétereket szolgáltat. Hátránya, hogy mivel általában maga a tRNS is *in vitro* kerül előállításra, nem tartalmazza a poszttranszkripciós módosításokat, melyek egyes esetekben funkcionális jelentőséggel bírnak. A módszer további hátránya, hogy meglehetősen munkaigényes, hiszen egyszerre csak egyféle enzim adott tRNS-sel való interakcióját vizsgálhatjuk.

Az *in vivo* rendszer (lásd az 1.3 ábrán) lényege az, hogy a tRNS-eket sejtekben állítjuk elő. A keletkezett tRNS-ekért ilyenkor mind a 20 aaRS enzim versenybe szállhat, és a funkcionális identitást az jellemzi, hogy az egyes enzimek egymáshoz képest milyen arányban fogadják el



1.2. ábra. Az identitáselem-meghatározás *in vivo* és *in vitro* módszereinek összehasonlítása.
Az ábrán egy fenilalanint szállító tRNS molekula egyes bázisait (identitáselemeit) lecserélve alanin identitásúvá alakítunk úgy, hogy a megfelelő pozíciókba az alanin identitáselemeinek megfelelő bázisokat építünk be.



1.3. ábra. Az identitáselem-meghatározás *in vivo* sémájának részletes bemutatása. Az 1.2 ábrával megegyezően Phe → Ala átalakítást mutatunk be.

szubsztrátként az adott tRNS-t. Mivel egy „nem normális" identitású tRNS sejtbéli megjelenése hibás aminosav-sorrendű fehérjék tömkelegét eredményezné, az ilyen tRNS-ek toxikusak a sejt számára. Ezért az ilyen *in vivo* rendszerekben csak olyan tRNS variánsokat lehet használni, melyek az antikodon hurokban egy stop kodont komplementáló tripletet hordoznak. Az ilyen tRNS-ek tehát nem szállítanak aminosavat normális kodonokhoz, ellenben egy adott stop kodont szupresszálnak. Az identitás meghatározásához egy rekombináns riporterfehérjét is termeltetnek a szupresszor tRNS mellett. A riporter fehérje mRNS-ének egyik kodon pozíciójában stop kodon van, és az izolált riporterfehérje szekvenálásán keresztül határozzák meg, hogy a szupresszor tRNS variáns milyen arányban szállította az adott kodonhoz az egyes aminosavakat a fehérjeszintézis során. A módszer lényeges előnyei az alábbiak: egyetlen kísérletben valójában 20 aaRS enzim interakciójáról kapunk képet, a poszttranszkripciós tRNS módosítások kialakulhatnak, és végül a módszer nemcsak a tRNS identitásáról, hanem általános használhatóságáról (pl. képes-e stabilan fennmaradni a sejtben és részt venni a transzlációban) is információt nyerünk. A módszer hátránya azonban az, hogy nem vizsgálhatók vele olyan tRNS típusok, melyek identitásában – tehát az aaRS enzimmel való speciális kapcsolatban - maga az antikodon hurok is meghatározó. Mint azt már említettem, a tRNS-ek javarésze hordoz identitáselemet az antikodonban.

1.5. Az in silico identitásvizsgálat lehetőségei

Az identitásvizsgálat elsődleges, kézenfekvő módszerei kísérletes jellegűek (lásd fent), azonban ezek költség- és időigénye igen magas. Az egyre több elérhető, sikeresen megszekvenált genom, valamint az ezekhez létrehozott szekvencia-adatbázisok ugyanakkor megnyitották annak elvi lehetőségét, hogy in silico módszerekkel határozzuk meg a tRNS-ek identitáselemeit, jelentős költség- és időmegtakarítást elérve ezzel. Jelen munka pontosan ezt tűzi ki célul, figyelembe véve azt, hogy a számítógépes és általában a predikciós módszerek kísérletesen meghatározott tényekből kell, hogy kiinduljanak, valamint elfogadva azt a tényt, hogy csak akkor válhatnak teljes értékűvé, ha a megfogalmazott állítások a laboratóriumban igazolást nyernek.

1.5.1. tRNS-ek in silico meghatározása genomi szekvenciákon

Amíg a tRNS-ekre vonatkozó térszerkezeti adatok még mindig rendkívül hiányosak, a genomszekvenálásoknak köszönhetően tRNS szekvencia-adatok özöne áll rendelkezésünkre. A genomi szekvenciák annotációjakor a tRNS szekvenciákat is azonosítani kell, azonban ehhez nem alkalmazhatóak a fehérjekódoló szakaszokra kifejlesztett predikciós módszerek, hanem külön,

speciális eljárások segítségével kell felderíteni. Itt, és a későbbi in silico irodalmi adatok megismerésében sokban hagyatkoztam David H. Ardell összefoglaló munkájára [46].
Jelenleg két elterjedt algoritmus szolgál a tRNS-ek genomi felderítésére. Az egyik a tRNAscan-SE [47], amely közel száz százalékos hatékonysággal, igen kis fals pozitív találati aránnyal működik. Az algoritmus az úgymond „kovariancia modelleket" [48] használja, amelyek a tRNS-ek általános szekvencia-hasonlóságait illetve a másodlagos szerkezet kialakításához szükséges törvényszerűségeket (bázispárosodások vagy éppen az egymással párosodni képtelen bázisok létét adott pozíciókban) veszi figyelembe. Ezt a keresőalgoritmust már odáig fejlesztették, hogy a keresés eredményéül nem ad hibás tRNS szekvenciát az emberi genomban. Kivételt ez alól csak néhány különlegesebb eset jelent, például az ugráló gének tRNS darabokat tartalmazó régiói, vagy a tRNS pszeudogének. A tRNS gének mindössze fél százalékát nem képes detektálni. Működtetésének csupán időkorlátja van.
A másik, igen hatékony eljárás, az ARAGORN [49] éppen a sebességet növeli azzal, hogy heurisztikus keresési modellt alkalmaz. További felhasználóbarát tulajdonsága, hogy jóval kevesebb paramétert kell meghatározni működéséhez: nem szükséges például megadni, hogy az adott szekvenciák milyen taxonómiai csoportba tartoznak. A szelektivitásban azonban valamelyest alulmarad a tRNAscan-SE-hez képest.

1.5.2. tRNS adatbázisok

A tRNS-ek genomi adatbázisokból történő kinyerése után alkalom nyílik a tRNS szekvenciák külön adatbázisba rendezéséhez. Az egyik legrégebben meglévő tRNS adatbázis Mathias Sprinzl munkájához fűződik. Ennek első verziói [50] még nem is támaszkodhattak a genomszekvenálások adataira, a tRNS szekvenciákat egyedi szekvenálások alapján vitték be, az egyes adatok annotációi pedig kézzel, szekvenciánként történtek meg. Ennél fogva ez az adatbázis mindamellett, hogy kevés rekordot tartalmaz igen megbízható. Például minden egyes szekvenciához irodalmi referencia is tartozik, illetve a tRNS-ek másodlagos szerkezete is jól ellenőrzött. Nem mellékes, hogy a tRNS-eknél használatos pozíciószámozást is ez az adatbázis honosította meg.
A Sprinzl-féle adatbázis folyamatos fejlődésen ment keresztül, legutolsó verzióján [51] alapul a tRNAdb adatbázis [52]. Az adatbázis nagy előnye, hogy a genomszekvenálásokból származó nagy mennyiségű adat mellett a szekvenciákhoz másodlagos szerkezeti adatokat is társít, illetve megőrzi a Sprinzl adatbázis jól bevett konvencióit, számozásai, annotációit. Az adatbázis további előnye, hogy keresési opciói nagyon jól paraméterezhetőek (például taxonok, taxon csoportok szerint, identitás szerint stb). Az adatbázis mindezek mellett a posztranszkripciós módosításokat is sok esetben tartalmazza. Szintén genomi adatokból épül föl a „*Genomic tRNA*

1.3. táblázat. Az IUPAC jelöléskonvenciója (Sebestyén Endre nyomán)

Szimbólum	Jelentés	Komplementer	Magyarázat
A	A	T vagy U	Adenin
C	C	G	Citozin
G	G	C	Guanin
T vagy U	T	A	Timin vagy Uracil
M	A vagy C	K	aMino
R	A vagy G	Y	puRin
W	A vagy T	W	Weak (gyenge; 2 H kötés)
S	C vagy G	S	Strong (erős; 3 H kötés)
Y	C vagy T	R	pYrimidine (pirimidin)
K	G vagy T	M	Keto
V	A, C vagy G	B	nem T vagy U
H	A, C vagy T	D	nem G
D	A, G vagy T	H	nem C
B	C, G vagy T	V	nem A
N vagy X	A, C, G vagy T	X vagy N	aNy (bármely)

Database" (GtRNAdb), amelyet már az említett tRNAscan-SE fejlesztői készítettek [53]. A szekvenciákhoz grafikus másodlagos szerkezetet is készít az adatbázis működtető motor. A genomi adatokból prediktált tRNS adatokhoz statisztikai elemzések, többszörös szekvencia-illesztések is elérhetőek. Az egyik legfrissebb adatbázis a tRNAdb-CE [54, 55] nagy előnye, hogy az automatikus annotációkat kézzel is ellenőrizték. A már ismertetett tRNAscan-SE és ARAGORN programok mellett a tRNAfinder nevű [56] eljárást is használja a genomi adatok elemzéséhez. Nagy előnye, hogy az ősbaktériumok egy különleges tRNS fajtáját, az ún. split-tRNS-eket [57] tartalmazó SPLITSdb [58] adatbázist is magába olvasztotta. A split-tRNSek olyan működőképes, teljes mértékben funkcionáló tRNS molekulák, amelyeknek két különböző darabja más-más génen kódolt. Felderítésük így a szokásos algoritmusokkal nem megoldható.

1.5.3. A nukleotidok és azok csoportjainak IUPAC jelöléskonvenciója

Az említett adatbázisok illetve a jelen dolgozat is sokszor használja a nukleotidokra, illetve azok csoportjaira a IUPAC (International Union of Pure and Applied Chemistry) jelöléseit. Ez azért praktikus, mert ahelyett, hogy 2-3 bázist felsorolnánk, egyetlen betűvel meg lehet jeleníteni azokat. A jelöléskonvenciót összefoglalóan a 1.3. táblázat mutatja be.

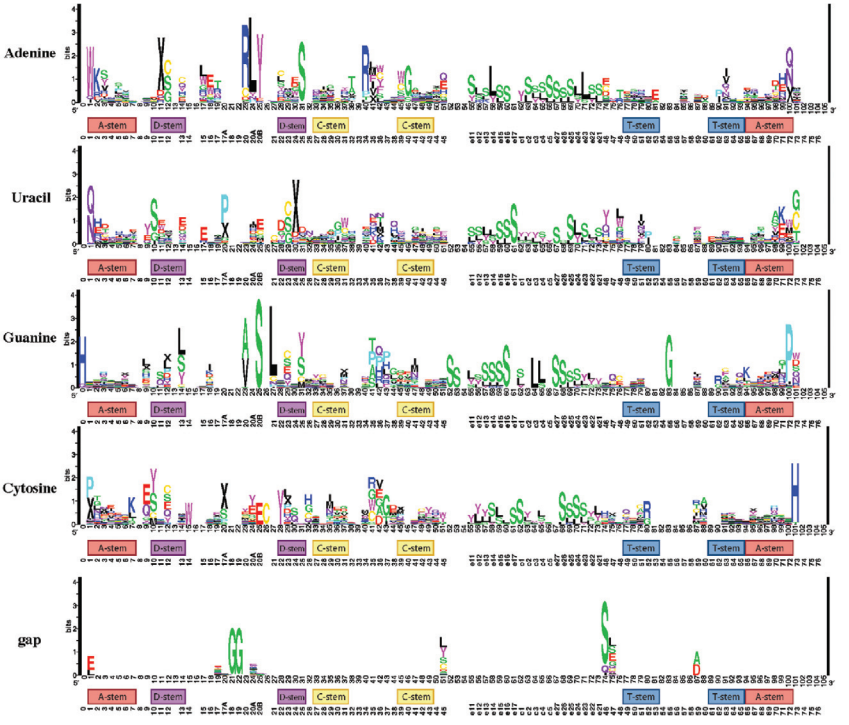
1.5.4. A szekvencia „*logo*”-k

Az úgynevezett szekvencia „*logo*”-k [59–61] felhasználása egyre elterjedtebb a különböző, többszörös illesztést bemutató eljárásokban. A módszer lényege az, hogy a szekvencia-információt a Shannon-entrópia [62] segítségével jeleníti meg. A többszörös illesztésben előforduló információnak, az ún „*sequence logo*”-k esetében (nukleinsavak esetében) nukleotidoknak valamilyen előfordulási valószínűségük (p) van. A szekvencia logo esetében egy-egy pozícióban az oszlopmagasság azzal arányos, hogy mennyire kevéssé random abban a pozícióban az adott elemek (esetünkben a bázisok) előfordulása. Maximálisan véletlenszerű (egyenletes) eloszlás esetében az oszlopmagasság nulla, minimálisan véletlenszerű esetben (amikor csak egyetlen bázisfajta van jelen) az oszlopmagasság maximális. Az oszlopon belül az egyes elemek relatív magassága a relatív gyakoriságukkal arányos. A többszörös illesztésben a szekvencia egyes pozíciói az ábrázolásban egymás mellé kerülnek, maguk az ábrázolt nukleotidok (vagy fehérjeszekvenciák esetében aminosavak) pedig egymás fölé: a legtöbbet előforduló legfölülre, alá pedig az arányaiban kevesebbszer szereplő elemek. Így könnyen vizualizálni tudunk egy nagy sereg szekvenciát, kiemelve a jellemző elemeket. A szekvencia logo tehát tulajdonképpen egy tömör vizualizálása a szekvencia-sereg pozíciónkénti információtartalmának.

1.5.4.1. A „*function logo*”-k alkalmazása tRNS molekulákra

Az információs-teória felhasználása is új lehetőségeket nyitott meg nem csak a tRNS szekvenciák vizualizációjában, hanem a determinánsok és antideterminánsok predikciójában is [2, 63]. Ehhez David H. Ardell és munkatársai egy új szempontú megközelítést vezettek be a „*sequence logo*”-k használatában. A tRNS-ek esetében a tRNS szekvenciákból nem az egyes pozíciók nukleotidjait ábrázolták, hanem azt, hogy ha az egyes pozíciókban valamelyik nukleotid vagy éppen „*gap*” szerepel, akkor ez a tény milyen mértékben kapcsolt az egyes tRNS-ek identitásához. Tehát azt ábrázolták „*logo*”-kal, hogy ha egy pozícióban például adenin szerepel, akkor az adott pozícióban adenint hordozó tRNS szekvenciák közül milyen gyakran fordulnak elő az egyes aminosav identitású tRNS molekulák. Ez az ábrázolás az ún. „*function logo*” tehát az egyes funkciókat (identitásokat) vizualizálja.

A fenti logikát könnyen megérthetjük a 1.4 ábra tanulmányozásával. Megfigyelhető, hogy az első pozícióban azok a szekvenciák, amelyek adenint tartalmaznak, főleg a triptofán (W) identitású tRNS molekulák közül kerülnek ki. Ugyanebben a pozícióban az uracil a glutamin és az aszparagin tRNS-eire jellemző (Q és N). Ardellék rávilágítottak arra is, hogy a „*function logo*”-k sok esetben a tRNS-ek identitáselemeit is kiemelik. Ilyen például a hisztidin G0 és C73 „*logo*”-i, amelyek ismert identitáselemek az *E. coli* tRNSHis-ben (az elemzést bakteriális adatokra készí-



1.4. ábra. tRNS-ek *„function logo”*-ja [2]

A *„function logo”*-k Sprinzl módosított, bakteriális, illesztett tDNS adatbázisából készült, a T-t U-nak ábrázolták. A *„logo”*-k sorszáma alatt a Sprinz-féle számozás is fel lett tüntetve. A *„logo”*-k azokat az aminosavspecifitású tRNS-ek hordozott aminosavának egybetűs kódját jelenítik meg, amelyek jellemzően az adott pozícióban az adott nukleotidot hordozzák. Részletes magyarázat a szövegben.

tették).

1.5.4.2. Az „*inverse logo*”-k

Az eddig tárgyalt „*logo*”-k a leggyakoribb elemeket hangsúlyozzák ki, ennél fogva nem látszik, hogy egy-egy adott elem ritka, vagy éppenséggel soha nem fordul elő. Ennek vizualizációjához vezették be az inverz „*logo*”-kat. A szekvencia „*logo*”-k logikáját megfordítva tehát nem csak azokat az információkat ábrázolhatjuk, amelyek a többszörös illesztésben jelen vannak, hanem azokat is, amelyek alulreprezentáltak, vagy éppen hiányoznak [64]. Ahogyan a „*sequence logo*”-ból, az „*inverse logo*”-ból is képezhetünk „*function logo*”-t, amely nem más, mint az „*inverse function logo*”. Ez azokat az identitásokat emeli ki, amelyekből a megfelelő bázisok hiányoznak az adott pozícióban.

1.5.4.3. A „*logo*”-k formalizálása

A különböző típusú „*logo*”- közül a mi esetünkben (tRNS-ek) a „*function logo*”, az érdekes, amelynél egy-egy aminosav logójának magasságát egy adott pozícióbanaz alábbi egyenlet segítségével számolhatjuk ki:

$$h_l(y|x) = \frac{\frac{p_l(y|x)}{p(y)}}{\sum\limits_{w\in\mathcal{Y}} \frac{p_l(w|x)}{p(y)}} I_l(\mathcal{Y}|x) \tag{1.1}$$

ahol az I a szekvencia információ (amely a Shannon-entrópiából vezethető le), az l az adott pozíció, a p a valószínűség, y az aminosav (az aminosavak halmaza Y), x és w pedig a bázisok (halmazuk az χ). A szekvencia információt pedig az alábbiak szerint számolhatjuk ki:

$$I_l(\mathcal{Y}|x) = H(\mathcal{Y}) - \mathrm{e}\Big(n(x)\Big) - H_l(\mathcal{Y}|x) \tag{1.2}$$

$$H(\mathcal{Y}) = -\sum_{y\in\mathcal{Y}} p(y)\log_2 p(y) \tag{1.3}$$

$$H_l(\mathcal{Y}|x) = -\sum_{y\in\mathcal{Y}} p_l(y|x)\log_2\Big(p_l(y|x)\Big) \tag{1.4}$$

ahol e egy korrekciós faktor, amely értéke annál nagyobb, minél kisebb elemszámú mintából indulunk ki, sok bemenő adat esetén értéke elhanyagolható [65].

1.6. A makromolekula-funkciók feltárásának általános elvei

A makromolekulákkal (fehérje, DNS, RNS) kapcsolatban leginkább elfogadott paradigma szerint a szekvencia meghatározza a makromolekula térszerkezeti, dinamikai és ezeken keresztül

funkcionális tulajdonságait. A molekuláris biológia egyik legnagyobb kihívása a szekvenciától a funkcionális tulajdonságokig vezető kapcsolatok algoritmikus leírása. Bár egy ilyen teljes, algoritmikus leírás (mely a molekuláris környezetet is figyelembe veszi) elméletileg lehetséges, egyelőre csak közelítő eredmények ismeretesek. A teljes megoldáshoz képest egy jóval szerényebb, de nagy gyakorlati jelentőségű lépés annak felismerése, hogy a szekvencia egyes elemei nem azonos szereppel bírnak a szerkezet és funkció kialakításában. Amennyiben vannak kitüntetetten fontos szerepű pozíciók, ezek felderítése nagyban leegyszerűsítheti a fenti probléma megoldását. Az ilyen kulcspozíciók feltárására jelenleg három fő megközelítés ismert. Az egyik esetében akár célzott kémiai módosításokkal, akár irányított mutagenezis segítségével lokális változásokat hoznak létre a szekvenciában, és ezeknek a változásoknak a szerkezeti és funkcionális hatásait vizsgálják genetikai, illetve biokémiai vizsgálatokkal. A másik megközelítés feltételezi a makromolekulák, illetve ezek komplexeinek (röntgen krisztallográfiával illetve NMR technikával megoldott) atomi felbontású térszerkezetének ismeretét. A szerkezet alapján általában kijelölhetők azok a pozíciók, melyek a térszerkezet létrehozása, illetve az ismert funkció szempontjából kiemelt fontosságúak lehetnek. A harmadik megközelítés azon alapul, hogy a szerkezet és a funkció konzervatívabb, mint a szekvencia, tehát a közös evolúciós eredetű makromolekulák szekvenciáiban a funkció szempontjából fontos pozíciók egymáshoz hasonlóbbak, mint a kevésbé fontos elemek. Amennyiben számos homológ szekvencia áll rendelkezésre, ezek statisztikai analízise kiemelheti az egyes kulcspozíciókat. A három említett megközelítés nem kizárólagos, sőt erősítik egymást. A mutációk hatásából, valamint a szekvencia analízisekből kapott eredmények sokszor csak a térszerkezet ismeretében válnak értelmezhetővé, míg a térszerkezet ismerete nagyban elősegíti a leginkább informatív genetikai-biokémiai vizsgálatok megtervezését.

A fenti három, ma már klasszikusnak számító megközelítés kombinálása rendkívül eredményesnek bizonyul, hiszen főleg ennek köszönhető a biológiai ismeretek soha nem látott ütemű gyarapodása. Ugyanakkor az is tény, hogy mindhárom megközelítés lényeges korlátokkal bír. A genetikai és biokémiai kísérletek egyik nagy problémája az, hogy a makromolekulák nagy mérete miatt a megvizsgálható mutációk, illetve ezek kombinációinak száma rendkívül magas, a kísérletek pedig rendszerint költségesek. A térszerkezeti vizsgálatokhoz vagy jó minőségű kristály kell (röntgen krisztallográfia), vagy nagy mennyiségű anyag (NMR), és ezek egyike sem triviális feltétel. Ráadásul az NMR esetén szigorú méretkorlátok is érvényesek. A szekvencia analizáló eljárások három jellegzetes esetben ütköznek problémába: i) ha kevés a rendelkezésre álló homológ szekvencia, ii) ha a szekvenciák túlságosan hasonlóak, iii) ha a szekvenciák túl heterogének. Egy további koncepcionális problémakör, mely valójában mindhárom megközelítést

érinti, abban rejlik, hogy vajon az egyes kulcspozíciók hatása egymástól függetlenül érvényesüle (additív modell), vagy ezek együttműködnek (kooperatív modell). Az eddigi megközelítések inkább az additív modellekre támaszkodnak.

2.

Célkitűzések

Doktori munkám az in silico szekvencia-analízis eszköztárából merít, ugyanakkor mind a kiindulási adatok meghatározásakor, mind pedig az eredmények értelmezésénél igyekszem kísérleti adatokra támaszkodni.

- Első kérdésünk a munka kezdetekor az volt, hogy egy új, Jakó Éena által fejlesztett [66] új, diszkrét matematikai megközelítésen alapuló osztályozó módszerrel megdönthetjük-e azt a régi „dogmát", miszerint a tRNS-ek szekvenciális alapon nem választhatóak ketté a nekik megfelelő, adott aminoacil-tRNS szintetáz osztályuk szerint.
- Kíváncsiak voltunk arra is, hogy erre más, Shannon-entrópián alapuló módszer is képes-e, egy olyan módszer, amely akár identitáselemek jóslására is alkalmas lehet.
- Végső célként pedig azt tűztem ki, hogy tovább növeljem saját eljárásaink hatékonyságát úgy, hogy esetleg új, eddig nem ismert identitáselemekre is javaslatokat tudjak tenni.

3.

Módszerek

3.1. Programok és programnyelvek

A biológiai kérdések megválaszolásának in silico eszközei közül jelen munka elsősorban a szekvencia-analízis tárából merített. A genomszekvenálások eredményeit, a már említett adatbázisok szekvencia-seregét a bioinformatika nem is olyan régmúltba nyúló „hőskorától" kezdve nagy, szöveges állományokban tárolják. Ezekben a szöveges állományokban a nyers szekvencia-adatokon túl a hozzájuk tartozó több-kevesebb információt is megtaláljuk (ún. annotációk formájában). Ezek az annotációk szintén szöveges információt jelentenek, a nyers szekvencia-adat mellett, attól elkülönülve, az adott adatbázis szabályrendszere szerint megállapított megkülönböztető jelzésekkel. Ezek a megkülönböztetések lehetnek egyszerű karakterek, rövidítések, szövegek, leggyakrabban valamilyen elválasztó karakterrel (szóközök megfelelő darabszámban, tabulátorok) az annotáció típusának jelzése és a hozzá tartozó információ között.

A szöveges állományok feldolgozásának legegyszerűbb és legkézenfekvőbb eszközei az ún „*script*" programozási nyelvek. A bioinformatika szekvencia-analízissel foglalkozó területén éppen ezért a Unix/Linux shell környezet a legalapvetőbb munkaeszköz (úgy is fogalmazhatnék, hogy a bioinformatikus pipettája).

A sokszor ismétlődő rutinfeladatokhoz számtalan esetben nyújt kiváló segítséget az EMBOSS programcsomag:

http://emboss.sourceforge.net/

A bonyolultabb, összetettebb feladatok végrehajtására Perl programnyelven írott „*script*"-eket, rövid programokat írtam, munkatársaim előszeretettel használták még a Python és Java nyelveket, amellyel készített programokat munkám során teszteltem és felhasználtam. A

statisztika problémák megoldásához, statisztikai eredmények megjelenítéséhez az erre egyik legalkalmasabb, szakterület-specifikus programozási nyelvet, az R-t választottam:
http://www.r-project.org/
A *„logo"*-k készítéséhez használt Makelogo program elérhető a Delila szoftvercsomagból (Schneider 1984):
http://www.ccrnp.ncifcrf.gov/ toms/delila.html
A tRNS-ek *„function logo"*-it a tRNALogofun programmal készítettem:
http://nar.oxfordjournals.org/content/suppl/2006/02/03/34.3.905.DC1/tRNAlogofun-1.0.zip

3.2. Felhasznált adatbázisok jellemzői

Az irodalmi áttekintésben részletesen bemutattam a tRNS adatbázisok típusait, előnyeit. Itt az egyes részfeladatokhoz felhasznált adatbázisokat sorolom föl.

3.2.1. A tRNomics adatbázis

A tRNS-ek szintetáz osztályaiknak megfelelő, szekvencia-alapú szétválasztásához Christian Marck és Henri Grosjean nagyszabású, jól annotált (csak valódi, működő tRNS géneket tartalmazó), megfelelően illesztett adatbázisát használtuk [67]. Azért volt ez az adatbázis ideális ehhez a munkához, mert a lehető legtöbb pontosan illesztett és ellenőrzött szekvenciát tartalmazta, és ebben az esetben nem alkalmaztam semmilyen szűrési lépést az adatbázisból kigyűjtött adatok esetében. A tDNS szekvenciák a három nagy *„kingdom"* (baktériumok, ősbaktériumok, eukarióták) adatait tartalmazták. Az adatbázist a szerzők bocsájtották rendelkezésünkre, akiknek ezúton is köszönetet kívánunk mondani. (Munkájuk címe révén a továbbiakban tRNomics adatbázis).
Az adatbázis 50 faj teljes tDNS készletéből áll, tehát az összes aminosavspecifitás minden izokakceptorát (az egy identitáshoz tartozó összes szekvencia-változatot) tartalmazza. A baktériumokból 30, az ősbaktériumokból 13, az eukariótákból pedig 7 faj adatait tartalmazza, a szekvenciák összes darabszáma 4204.

3.2.2. Az MSDB adatbázis

Az osztályspecifikus *„logo"*-k készítéséhez ugyanazt a módosított Sprinzl adatbázist (továbbiakban: MSDB - *„Modified Sprinzl Database"*) használtam, amelyet a *„function logo"* szer-

zői, hogy az eredmények könnyen összehasonlíthatóak legyenek. Az adatbázis elérhető innen: http://nar.oxfordjournals.org/content/suppl/2006/02/03/34.3.905.DC1/MSDB.aln.txt Az adatbázis összesen 655, kizárólag bakteriális, nem redundáns (tehát a szekvenciálisan teljesen azonos izoakceptoroktól megtisztított) tDNS szekvenciát tartalmaz.

3.2.3. A tDNAdbC adatbázis

Az eddig nem ismert, potenciális aminosav-identitáselemek jóslásához a tDNS szekvenciákat a tRNAdb adatbázisból [52] töltöttem le a baktériumok és az eukarióták estében.
Az adatbázis elérhető online: http://trnadb.bioinf.uni-leipzig.de/ Az ősbakteriális szekvenciákat a tRNAdb-CE adatbzisból vettem (http://trna.nagahama-i-bio.ac.jp) ugyanis jelenleg csak ez tartalmazza a SPLIT tRNS-eket. (Az említett két adatbázisból származó adatokat összefoglalóan a dolgozatban tDNAdbC-nek nevezem: „*Complex*" tDNS adatbázis). Ennél az analízisnél a nyers szekvencia-adatokat bizonyos szempontok alapján szűrtem (lásd a „*Módszerfejlesztés*" című fejezetben).

3.3. Az ECP algoritmus

Az „*Extended Consensus Partition*" (továbbiakban: ECP) eljárást Jakó Éena dolgozta ki [66]. Az eljárás alkalmazható bármilyen két nukleotid-szekvencia csoportra, de akár fehérjékre is. Munkám során és a dolgozatban tRNS/tDNS szekvenciákon alkalmaztam.

3.3.1. Az SCP algoritmus

Az ECP ismertetése előtt a „*Strict Consensus Partition*" (SCP) működését is bemutatom, ugyanis az ECP és az SCP összehasonlítására is sor kerül majd (lásd később), illetve az ECP-vel elérhető eredményeket az SCP-hez mérten fogom értékelni. Az SCP képzése során mindig illesztett szekvenciákból indulunk ki. A konszenzust az illesztett szekvenciák adott pozícióiban képezzük. Az, hogy az SCP „*strict*", azaz szigorú, azt jelenti, hogy amennyiben az illesztett szekvenciák adott pozíciójában minden egyes bázis ugyanaz, akkor annak „*SCP*"-je nem más, mint az adott bázis. Konvencionális szóhasználattal élve az adott pozícióban ez a szekvenciasereg konszenzusa. A későbbiekben ezt a konszenzust, tehát az adott pozícióban található, minden egyes szekvenciában megegyező nukleotidot „*strictly present*" (SP) elemeknek nevezzük. A szigorú („*strictly*") jelleg nem mást jelent, mint azt, hogy itt minden egyes szekvencia számít, tehát a konszenzust eltörölheti akár egyetlen, addig még nem szereplő bázis megjelenése valamelyik

szekvenciában a többszörös illesztés vizsgált pozíciójában. Ha az SCP-vel két többszörösen illesztett szekvenciasereget kívánunk összehasonlítani, akkor a különbségeket csak az egyik illetve másik szekvenciaseregre képzett SP elemekkel írjuk le.

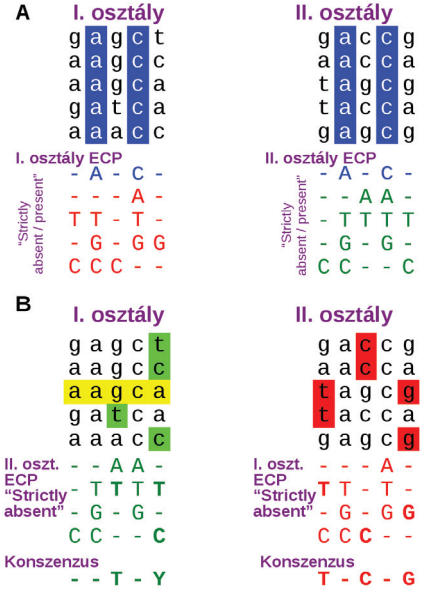
3.3.2. Az ECP rövid, mesterséges szekvenciákon

Az SP elemek mellett a többszörösen illesztett szekvenciákban meghatározhatjuk a *„strictly absent"* (SA) elemeket is. Ez azt jelenti, hogy pozíciónként számba vesszük azokat a bázisokat, amelyeket egy-egy adott pozícióban egyetlen egy szekvencia sem tartalmaz. Az elemek *„strict"* jellege tehát itt is ugyanazt jelenti, mint az SP elemeknél: egyetlen eltérő elem felülírhatja a pozíció jellegét. Ebből adódik a módszer érzékenysége. Az SP illetve SA elemek képzését a 3.1 A ábra mutatja be.

Az SCP-hez hasonlóan az ECP-vel is össze tudunk hasonlítani két, többszörösen illesztett szekvenciasereget. Az egyik többszörösen illesztett szekvenciasereget nevezzük I. osztálynak, a másikat pedig II. osztálynak (rövid szekvenciás példát lásd 3.1 ábra). Először pozíciónként meghatározzuk a I. osztály SA elemeit, majd a másik, II. osztályban szintén pozíciónként megvizsgáljuk azt, hogy ugyanabban a pozícióban tartalmaz-e valamelyik szekvencia olyan elemet, ami *„strictly absent"* az I. osztályban. Ha a II. osztály bármelyik szekvenciája az adott pozícióban tartalmaz olyan bázist, amely *„strictly absent"* az I. osztályban, akkor az a bázis, vagy azok a bázisok lesznek a II. osztály ún. „diszkrimináló elemei" (DE), hiszen ezek elkülönítik a II. osztályt az I-től. Amennyiben egy-egy pozícióban egynél többféle bázist kapunk ezzel a módszerrel, mint DE, abban az esetben a bázis-csoport megfelelő egybetűs IUPAC kódját használjuk. Ez után, ugyanilyen módon meghatározzuk a I. osztály DE-it a II. osztály SA elemeinek segítségével (lásd 3.1).

Ha egy vizsgált osztályban egy adott szekvencia valamelyik pozícióban tartalmaz a másik osztály ugyanazon pozíciójában található SA elemek közül legalább egyet, akkor ebből a másik osztályból „kizárja" saját magát. Ugyanakkor azt a szekvenciát, amelyik egyetlen egy pozíciója sem tartalmaz ilyen kizáró elemet (DE-t) egyetlen osztályból sem, *„fals pozitív"* szekvenciának nevezzük, hiszen úgy szerepel az adott osztályban, hogy ezek alapján a szabályok alapján akár a másikban is szerepelhetne. Más szóval az ilyen szekvenciák „nem szólnak bele" a DE-k képzésébe, bármelyik osztályba is sorolnánk őket a DE-k előállításához.

Az ECP módszer és az általa definiált szekvencia-távolságok megértéséhez vegyünk még egy egyszerű példát. Az osztályozáskor, két adathalmaz (szekvenciasereg) egymástól való elkülönítéséhez képezzük a diszkrimináló elemeket. Két adathalmaz akkor áll egymáshoz legközelebb, ha – most csak egy pozíciót vizsgálva – ha mindkét adathalmaz ugyanazt a bázist tartalmazza:



3.1. ábra. Az ECP működése rövid, mesterséges szekvenciákon

A) Két, többszörös illesztésből indulunk ki (I. és II. osztály), amely mesterséges szekvenciákat tartalmaz. A osztályok SP elemeit az eredeti szekvenciákon kék háttérrel illetve kék színnel, az I. osztály SA elemei pirossal, a II. osztály SA elemei pedig zölddel jelöltek.

B) Az ECP diszkrimináló elemeit (DE, magenta színnel jelölve) az I. osztályra úgy kapjuk meg, hogy a II. osztály SA elemeit az adott pozícióban kijelöljük (zöld háttérrel kiemelve az I. osztályba tartozó szekvenciák között illetve a II. osztály SA elemei közül vastag betűvel) és – amennyiben több egyezést is találunk – IUPAC konszenzusát képezzük. Utóbbira példa az utolsó pozíció Y-ja. Ugyanezt az analízist a II. osztályon is bemutatja az ábra, itt az I. osztály SA elemei pirosak, az A) ábrarészlettel megegyezően.

Azt az szekvenciát, amelyik egyetlen egy pozíciója sem tartalmaz DE-t, *„fals pozitív”*-nak tekintjük, az ábrán sárga háttérrel szerepel [66].

nem találunk DE-t. Az ECP már megismert jellegéből adódóan ugyanakkor a DE-k száma szintén nulla lesz, ha mindkét szekvencia-halmaz ebben a pozícióban mind a négy bázist megengedi. Ezt a két, egymástól szekvenciális alapon a lehető legkülönbözőbb esetet az ECP nem különbözteti meg. (Ennek okait és következményeit lásd majd a *„Konklúzió"* részben.)
A diszkrimináló elemek tehát arra jók, hogy két vizsgált szekvenciacsoportot (osztályt) egyértelműen el tudjunk különíteni egymástól: az ECP nem egyedi szekvenciákat, hanem azok halmazait, a halmazok egymástól való távolságát adja meg a szekvenciatérben.

3.4. Statisztikai módszerek

3.4.1. Az ECP hatékonyságának tesztelése

Ahhoz, hogy megvizsgáljuk, hogy az ECP módszer milyen hatékonyan választja szét a két osztályt, illetve ahhoz, hogy a diszkrimináló elemek egyediségét megállapítsuk, három különböző statisztikai módszert vezetettünk be.

3.4.1.1. Az osztály-szétválasztás hatékonysága

Ahhoz, hogy az ECP módszer jóságát teszteljük, összehasonlítottuk a korábban alkalmazott SCP módszerrel. Arra voltunk kíváncsiak, hogy az ECP az SCP-nél hatékonyabban tudja-e szétválasztani a tDNS szekvenciákat a nekik megfelelő szintetáz osztályba tartozásuk szerint. Ehhez Ittzés Péter írt algoritmust és egy programot (publikálatlan eredmény), amellyel *„bootstrap"* analízist tudtunk végezni. Ebben az analízisben a tDNS szekvenciákat véletlenszerűen osztottuk be két osztályba, amelyek mérete az eredeti I. és II. osztály (*„a priori"*) méretével egyezett meg. Az összes lehetséges, véletlenszerűen előállítható két osztályt létrehoztuk. Ezután a véletlenszerűen generált osztályokra mind az SCP, mind az ECP analízist elvégeztük és feljegyeztük a *„fals pozitív"* szekvenciák számát, tehát azoknak a szekvenciáknak a számát, amelyek mindkét osztályba tartozhatnának. Ezt a számot összehasonlítottuk az eredeti osztályok (a szintetázuknak megfelelően beosztott tDNS szekvenciák) *„fals pozitív"* szekvenciáinak számával. Belátható, hogy minél jobban szeparálódik szekvencia tulajdonságok szerint két osztály, annál kevesebb lesz a *„fals pozitív"* szekvenciák száma. Akkor tekintettük szignifikánsnak az eredeti osztályokra kapott eredményt, ha az összes előállt esetből 25%, vagy annál kevesebb esetben volt a véletlenszerűen előállított osztályok analízisből kapott *„fals pozitív"* szekvenciáinak száma kevesebb az a priori, tehát valós osztályból képzett *„fals pozitív"* szekvenciák számánál.

3.4.1.2. Az osztályra jellemző DE-készletek egyedisége

Az osztály-szétválasztás hatékonyságához hasonlóan megvizsgáltuk az egyes, *„a priori"* osztályok DE elemeinek egyediségét. Ezt a fent vázolt *„bootstrap"* analízishez hasonlóan tettük meg Kun Ádám munkájának nyomán, azzal a különbséggel, hogy itt nem tartottuk meg az eredeti osztályméreteket, hanem ugyanakkora méretű mesterséges osztályokat állítottunk elő. Ebben az analízisben azt vizsgáltuk meg, hogy a véletlenszerűen létrehozott osztályok ugyanazt a DE-készletet hozzák-e létre, mint az *„a priori"* osztályok.

3.4.1.3. Az egyes DE-k egyedisége

A fentieken túl megvizsgáltuk azt is, hogy az egyes fajokban vannak-e olyan pozíciók, amelyben az egyes osztályokhoz tartozó DE-k egyediek. Tehát az osztály-szétválasztás egyediségénél leírtaknak megfelelően véletlenszerűen osztályokat képeztünk, és azokat a diszkrimináló elemeket gyűjtöttük ki, amelyek az *„a priori"* osztályban szerepeltek és a véletlenszerűen generált osztályokban csak az esetek maximum 5%-ában jelentek meg.

3.4.1.4. Az aminosavidentitásokra alkalmazott ECP statisztikai elemzése

Az elemzéséhez (lásd még: *„Módszerfejlesztés"* című fejezetben) Pearson illetve Spearman korrelációs analízist használtam, valamint Pál Gábor fejlesztett egy speciális *„bootstrap"* típusú analízist. A korrelációkat, valamint a *„bootstrap"* módszer pontos leírását és az eredményeket az *„Eredmények és értelmezésük"* című fejezet alatt mutatom be.

4.

Módszerfejlesztés

4.1. Az adatbázisok átalakítása; saját, szűrt adatbázisok készítése

A munkám során az egyes adatbázisokon különböző módosításokat kellett végeznem ahhoz, hogy megfelelő bemenő adathalmazt szolgáltassanak az egyes vizsgálatokhoz. Erre elsősorban az analízis különböző módszereinek érzékenysége illetve egyes tRNS/tDNS szekvenciák sajátosságai miatt volt szükség. Jelen fejezet pusztán arra szorítkozik, hogy az átalakítási, illetve szűrési lépéseket bemutassa. Amennyiben a szűrés befolyásolhatja vagy befolyásolta valamilyen értelemben az eredményeket, úgy azt az *„Eredmények és értelmezésük"* részben külön kiemelem.

4.1.1. A tRNomics feldolgozása

A tRNomics adatbázis feldolgozásakor a legfőbb feladatot a tDNSLys szekvenciák besorolása jelentette. Ahogyan a bevezetőben is bemutattam, a LysRS-ek két különböző osztályba is tartozhatnak. Az UniProtKB-SwissProt domén adatbázisából letöltöttem a megfelelő szintetáz enzimhez tartozó rekordokat, és az annotációk alapján különválasztottam az első illetve második osztályba tartozó szintetázokat. Ez alapján az egyes fajokat különválasztottam aszerint, hogy a Lys szintetáza(i) mely osztályba tartoznak. Az irodalmi adatoknak megfelelően [68] az eukarióta fajok Lys szintetázai mind, a bakteriális fajoké pedig javarészt a második osztályba tartoznak. A legtöbb vizsgált ősbaktérium – szintén az eddigi ismereteknek megfelelően – Lys szintetáza az első osztályba tartozik [17, 68, 69].
Több fajról azonban nem volt az adatbázisban domén-annotáció (Pyrobaculum aerophilum, Sulfolobus tokodaii, Ferroplasma acidarmanus, Sinorhyzobium meliloti). Ezen fajok szintetá-

zainak aminosavszekvenciáit kigyűjtöttem, és ClustalW program [70, 71] segítségével az összes első illetve második osztályba tartozó szekvenciával többszörösen illesztettem. A kapott dendogram segítségével megállapítottam a homológiákat.
Miután minden szekvencia osztályok szerinti hovatartozása egyértelmű lett, a tDNS-eket két külön csoportba, osztályuknak megfelelően rendeztem.
Végül a szekvenciákból a későbbi analízisekhez eltávolítottam a 0. pozíciót illetve a variábilis hurkot.

4.1.2. Az MSDB feldolgozása

Az MSD adatbázis feldolgozását nehezítette, hogy mindössze egyetlen annotációval látták el a szekvenciákat: csak az antikodon tripletjét és egy egyedi azonosító (sor)számot tettek közzé a szerzők. Noha szekvencia alapon jó eséllyel (pl. BLAST futtatásokkal) beazonosíthatóak lennének a kiindulási szekvenciák (bár – főleg mivel bakteriális szekvenciákról lévén szó – gyakoriak az akár fajok között redundáns szekvenciák), munkám szempontjából ez nem volt lényeges. Az osztályok szerinti bontást az antikodon tripletje alapján tudtam elvégezni.
A publikált adatbázison ezért nem változtattam, azonban készítettem egy, az adatbázist feldolgozó Perl scriptet (elérhető az online anyagok között), amely segítségével egy úgynevezett *„profile matrix”* készíthető. Ez a mátrix az egyes pozíciókban található négy nukleotid (a tDNS-ek esetében A, T, G és C) valamint a *„gap”*-ek gyakoriságát (konkrét darabszámát) mutatja meg. Ez a bemeneti állomány szükséges a TRNALOGOFUN program számára a *„logo”*-k kirajzolásához.

4.1.3. A tDNAdbC szűrése

4.1.3.1. Első szűrési lépés mindhárom adatcsoportra

Az első szűrési szempontot az egy-egy kingdomra megállapítható törvényszerűségek jelentették, amelyeket C. Marck és H. Grosjean tRNomikai elemzésükben állapítottak meg [67]. A szűrés azt jelentette, hogy azokat a tDNS szekvenciákat, amelyek az adott „szabályoknak” nem feleltek meg (az adott pozícióban nem a felsorolt nukleotidot vagy nukleotidok egyikét tartalmazták),eltávolítottam az adatbázisból.
A szabályok bakteriális szekvenciáknál: H14, G18, R19, Y33, G53:C61, T54, T55, Y56, D57, A58. Az eukarióta szekvenciáknál: Y8, Y11, A14, -17a, G18, G19, R21, R24, H32, Y33, R37, H38, G53, H54, T55, C56, R57, A58, C61 (a IUPAC jelöléseket használva). Mivel az ősbakteriális szekvenciákat (ide értve a *„split”* tRNS-eket is) tartalmazó tRNADB-CE adatbázis nem

tartalmaz illesztett tDNS szekvenciákat, ebben az esetben magam végeztem el az illesztést. A letöltött szekvenciákat Fujishima és munkatársai által közölt módszerrel (Fujishima 2009) illesztettem ClustalW programmal illetve manuálisan korrigáltam is az illesztést. Az elemzésből a variábilis hurkot az illesztés nehézségei és az eredmények szempontjából várhatóan kisebb jelentősége miatt kihagytam.

Az első szűrési lépésnél figyelembe vett szabályok az ősbakteriális szekvenciáknál a következők voltak (akár Fujishima-nál): Y8, A14, G15, G18, G19, R21, T33, Y48, G53, T54, T55, C56, R57, A58. A továbbiakban mindhárom *„kingdom"* esetében csak azokkal a szekvenciákkal dolgoztam tovább, amelyek a fenti kritériumoknak megfelelnek, tehát az adott pozíciókban a feltüntetett nukleotido(ka)t tartalmazzák. Így egyszerűen ki tudtam küszöbölni az esetleges adatbázis-hibákat, illetve az esetleges extrém, kérdéses funkcionalitású különleges tDNS szekvenciákat.

4.1.3.2. Második szűrési lépés a bakteriális és az eukarióta adatcsoportra

Második szempontom az adatbázis további szűrésekor, a kiindulási adatok előállításához az volt, hogy minden egyes identitás esetén csak olyan szekvenciákkal dolgozzam tovább, amelyek tartalmazzák az adott identitásra vonatkozóan már publikált [24] identitáselemeket (1. táblázat). Erre azért volt szükség, mert alkalmazott módszerünk „szigorú" jellegéből adódóan rendkívül érzékeny egy-egy, akár egyetlen nukleotid pozícióban eltérő szekvencia megjelenésére (lásd a *„Módszerek"* fejezetben írtakat). Célunkat, hogy a már meglévő identitás-elemek ismeretében a tRNS molekulák egyéb pozícióiban feltérképezzük az esetleg meglévő, még ismeretlen identitáselemeket, így nagyobb eséllyel tudjuk érni, mivel egy-egy kivételes, nem jellemző szekvencia megjelenése az összképet nem zavarja meg.

Felhívom viszont arra a figyelmet, hogy az egyes *„kingdom"*-ok esetén a már jól ismert modellfajok identitáselemei alapján szűrtem ki a szekvenciákat a kiindulási adatbázisból. A megállapított törvényszerűségek, egy-egy aminosav-specifitásra vonatkozó identitás elemek a baktériumok esetén az *E. coli*-ból, az eukariótáknál az élesztőből származnak. Az ősbakteriális szekvenciákat a kevés kísérletesen meghatározott identitáselem miatt nem szűrtem. Ennek megfelelően a szűrések elvégzése után csak az említett fajok rokon-szekvenciái maradtak az adathalmazban, a későbbi elemzések és a levont következtetések is csak a megfelelően szűkített faj-csoportra vonatkoznak.

A publikált identitáselemek közül csak a determinánsokat vettem figyelembe. Egy esetben, a baktériumoknál feltüntetett G15:G48 párt (Giegé 1998) kihagytam a szűrés feltételei közül, ez a tulajdonság ugyanis csak a gamma-proteobaktériumok tRNACys-ére jellemző, nagyon kevés

fajban található meg, a Haemophilus influenzae és az *E. coli* közös ősében jelenhetett meg

4.1.3.3. Harmadik szűrési lépés

Végül mindegyik adatcsoport esetében eltávolítottam a redundáns (teljesen megegyező) szekvenciákat, hogy csak egyedi tDNS-eket tartalmazzon adatbázisunk (az MSDB-hez hasonlóan).

Az 1. számú mellékletben feltüntettem a tDNAdbC adatbázisban szereplő fajokat, valamint azt, hogy a kiindulási adatokban, majd az egyes szűrési lépések után hány szekvenciát tartalmaz az adatbázis (rendre a bakteriális, eukarióta és ősbakteriális csoportokban).

A letölthető anyagok között elérhető a három szűrési lépés után létrejött adatbázis, szekvenciákkal és eredeti azonosítójukkal/annotációval multifasta formátumban mindhárom *„kingdom"* esetében. Szintén elérhetőek ugyanabban a fájlban a redundáns szekvenciák is a saját, eredeti azonosítójukkal/annotációjukkal.

4.2. Az ECP használata tRNS-identitásokra

Az ECP-t a korábbi leírás két osztályra vezette be[66], de valójában nem csak két osztály szétválasztására használhatjuk. A 20 tRNS-identitás készlet szétválasztására egy új eljárást fejlesztettem ki. Az alapelveken, a módszer algoritmusán nem változtattam, az elemzéshez azonban nem a két – I-es és II-es – aminoacil-tRNS szintetáz osztálynak megfelelő tDNS szekvenciákból álló csoportokat hasonlítottam össze (lásd korábbi munka), tehát nem ezek esetére határoztam meg a diszkrimináló elemeket, hanem aminosavspecifitás alapján képeztem szekvenciacsoport párokat. Az analízis során minden egyes aminosavspecifitású tDNS szekvenciacsoportot minden egyes, tőle különböző aminosavspecifitású tDNS szekvenciacsoporttal összehasonlítottam az ECP módszer segítségével, és meghatároztam azokat a diszkrimináló elemeket, amelyek arra az aminosavspecifitás csoportra jellemzőek. Így egy-egy fajcsoporton (baktérium, ősbaktérium és eukarióta) belül összesen 380 párt (a 20 aminosav a 19 másikkal szemben nem szimmetrikus módon) képeztem. A párokban egy fajcsoporton belül, egy aminosavspecifitáshoz tartozó, a fent említett szűrési módszerek után fennmaradt összes tDNS szekvencia van. A bakteriális és eukarióta szekvenciák esetében az adatbázisok alapján [52] az ősbaktériumoknál pedig a saját illesztésem szerint a szekvenciákon pozíciónként haladtam végig. Az adott pozícióban összegyűjtöttem a *„strictly absent"* (hiányzó) elemeket, amelyek tehát az adott identitáscsoport adott pozíciójában egyetlen szekvenciában sem fordulnak elő. Ezután identitás-páronként megvizs-

gáltam, hogy az eltérő identitású tRNS készletben az illesztett szekvenciák között a vizsgált pozícióban megtalálható-e a másik csoportból hiányzó *„strictly absent"* elem. Ha igen, akkor ez(ek) az elem(ek) a diszkrimináló elem(ek). A 380 pár elemzése során a párokra jellemző diszkrimináló elemeket gyűjtöttem össze.

4.2.1. Az AEV

Az átlagos kizárási értéket (*„average excluding value"* – AEV) azért vezettem be, hogy minden pozícióra külön-külön megállapíthassam, hogy abban milyen gyakorisággal fordulnak elő diszkrimináló elemek. Tehát minden egyes pozícióban minden identitásra megvizsgáltam, hogy a többi, eltérő másik identitás közül hány darab tartalmaz diszkrimináló elemet: így minden pozícióra kaptam egy összesített diszkrimináló elemszámot.
Az átlagértéket egy-egy pozícióban úgy állapítottam meg, hogy az adott pozícióban azonosított összes diszkrimináló elem számát elosztottam hússzal, vagyis az aminosavspecifitások számával. Az algoritmus működését rövid, mesterséges szekvenciákon az 4.1 ábrán mutatom be.

4.2.2. Az ECP módszer és az AEV formalizálása

Az AEV érték matematikai formalizálásához bevezetjük az $\mathcal{Y}$ változót. Az $\mathcal{Y}$ elemei nukleotidbázisok, tehát $\mathcal{Y} \in \chi$ ahol $\chi = \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$. A változó egyes $\mathcal{Y}_{ik}^{j}$ állapota nem más, mint az a bázis, amelyet egy adott i aminosav-identitás ($i = 1 \ldots, N, N = 20$) j-edik pozíciójában ($j = 1, \ldots, L, L = 96$ – o-tól a 73-as pozícióig) az identitáshoz tartozó k-adik szekvencia ($k = 1 \ldots, M_i$) tartalmaz. M_i fajonként és aminosav-identitásonként változik.

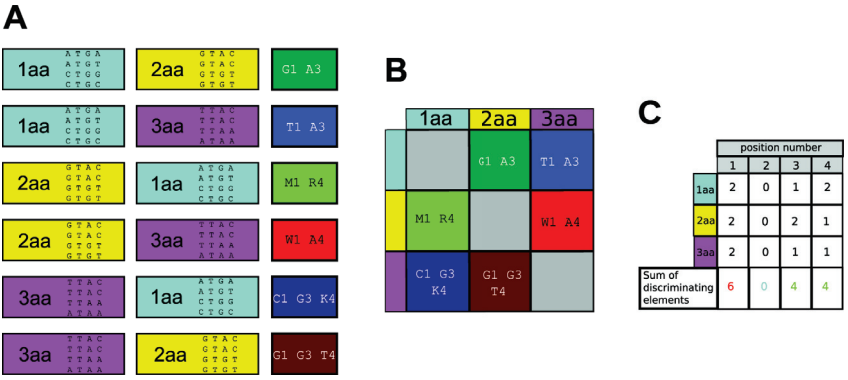
Bevezethetjük tehát azon bázisok halmazát, amelyek egy i identitás j-edik pozíciójában találhatóak:

$$\mathcal{Y}_i^j := \{\mathcal{Y}_{ik}^j | k = 1 \ldots, M_i\} \tag{4.1}$$

Az i aminosav-identitás diszkrimináló elemei (DE) az l aminosavval szemben (szintén: $l = 1 \ldots, N, N = 20$) a j pozícióban:

$$A_{il}^j := \left(\chi \setminus \mathcal{Y}_i^j\right) \cap \mathcal{Y}_l^j \tag{4.2}$$

Az „átlagos kizárási érték" (AEV) számításához a vizsgált j pozícióban defniáljuk az alábbi függvényt:



4.1. ábra. Az átlagos kizárási érték számítása rövid, mesterséges szekvenciákon.

A) Az ábrán három, mesterséges adatcsoportból, kvázi aminosavspecifitásból álló adattömegből indulunk ki. A három „aminosav" 1aa, 2aa és 3aa (ciánkék, sárga és magenta színű háttérrel). Minden „aminosavspecifitás" négy-négy szekvenciát tartalmaz (a téglalapokba írva). Minden szekvenciacsoportot párosítunk mindegyik másik csoporttal, az ábrán fentről lefelé rendre: 1aa-2aa, 1aa-3aa, 2aa-1aa, 2aa-3aa, 3aa-1aa, 3aa-2aa. Az egyes párok esetén meghatározzuk a diszkrimináló elemeket (eredmény a kisebb téglalapokban). A diszkrimináló elemek pl. az 3aa-2aa és az 2aa-3aa esetén a sötétpiros és piros hátterű négyzetekben vannak feltüntetve. Tehát nem mindegy, hogy 3aa-2aa vagy 2aa-3aa párt nézünk. Például az 1. pozícióban a 3aa identitás esetén hiányzó bázisok a G és C (SA elem). Ezek közül a 2aa szekvenciák tartalmaznak G-t, így diszkrimináló elemnek a G-t tekintjük (sötétpiros háttérben a G1). A 2aa szekvenciák az első pozícióban nincsen A, T és C, a 3aa identitás szekvenciái ugyanakkor ezek közül csak A-t és T-t (IUPAC kóddal W) tartalmaznak, tehát ezek hiánya a 3aa identitásban az, ami ténylegesen kizárják azokat a szekvenciákat, amelyek tartalmaznának ilyen bázisokat: ezek tehát a diszkrimináló elemek (DE).

B) A diszkrimináló elemek rövid, mesterséges szekvenciákon

A számított diszkrimináló elemek összefoglalását mutatja be az ábra. Az ábrán látható, hogy az egyes identitások nem szimmetrikus módon tudják kizárni egymást. Az A) ábrarésznél leírt példánál maradva a piros négyzetben azok a diszkrimináló elemek szerepelnek, amelyek a 2aa identitás egyes pozícióihoz tartoznak és a 3aa identitás szekvenciáit zárják ki. És vice versa: a sötétpiros négyzetben a 3aa identitás diszkrimináló elemeit találjuk, amelyek a már ismertetett okokból kifolyólag lehetnek más bázis(ok) más pozíció(k)ban.

C) A diszkrimináló elemek gyakorisága pozíciónként

Az algoritmus következő lépésében minden egyes pozícióra összegezzük a diszkrimináló elemek számát. Megnézzük minden egyes identitásra, hogy az adott pozícióban hány másik identitás hordoz diszkrimináló elemeket. Például a 3aa identitást a 3. pozícióban csak az egyik – az 1aa identitás – zárja ki. Ezután az egyes pozíciókban található diszkrimináló elemeket összegezzük (majd pedig elosztjuk az identitások számával – ez lesz az AEV érték, ami nem szerepel az ábrán: lásd a szövegben).

$$\mathbb{R}\left(A_{il}^{j}\right) := \begin{cases} 1, & \text{ha } A_{il}^{j} \neq \varnothing \\ 0, & \text{ha } A_{il}^{j} = \varnothing \end{cases} \tag{4.3}$$

Végül a függvény kapott értékeit minden identitás esetén minden identitással szemben összegezzük (tehát összeadjuk azokat az eseteket, amikor találtunk diszkrimináló elemeket a pozícióban), illetve elosztjuk az aminosavak számával:

$$n^{j} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{i=1 \\ i \neq l}}^{N} \mathbb{R}\left(A_{il}^{j}\right) \tag{4.4}$$

amely érték nem más, mint az AEV.

5.

Eredmények és értelmezésük

5.1. A tRNS szekvenciák szekvencia alapú szétválasztása szintetáz osztályuknak megfelelően ECP módszerrel

5.1.1. Az ECP tRNS/tDNS szekvenciákon

A „*Módszerek*" fejezetben leírtaknak megfelelően az ECP DE elemeit a rövid szekvenciákon bemutatott (3.1 ábra) módon határoztam meg valós, tRNS – estünkben tDNS – szekvenciákon. Az elemzett 50 faj közül az élesztő példáját mutatom be az 5.1 ábrán[66].

Egy-egy fajban tehát előállítható a két tRNS osztályra meghatározott DE-készlet. Ez azt jelenti, hogy azok a szekvenciák, amelyek a megfelelő pozícióban tartalmaznak a másik osztályból ugyanabban a pozícióban minden szekvenciából hiányzó elemet, azok „kizárják magukat" az ellentétes osztályból.

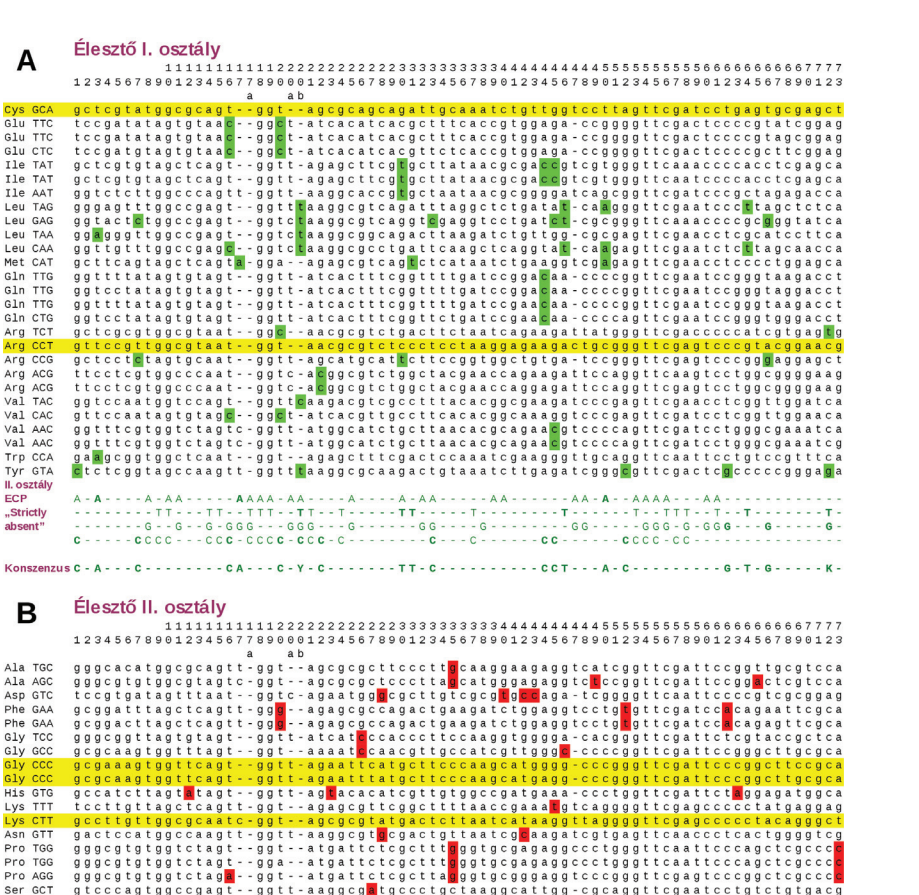
Ugyanezt az osztályzást mutatom be a konvencionális lóhere alakú kétdimenziós tRNS ábrázo-

---

5.1. ábra *(lásd túloldalt).* Az ECP algoritmus működése az élesztő tDNS szekvenciáin

Az ábra az élszesztő (*Saccharomyces cerevisiae*) összes tDNS szekvenciáját mutatja. Az A) ábrán a I. osztály a B) ábrán pedig a II. osztályú szintetázoknak megfelelő identitású tRNS géneket soroltam fel. Az A) ábra alatt a II. osztály osztály SA („*strictly absent*") elemei vannak feltüntetve, zöld színnel. Vastagon kiemeltem azokat a bázisokat, amelyeket az I. osztály szekvenciái közül az adott pozícióban legalább egy tDNS szekvencia tartalmaz Azt az I. osztályba tartozó tDNS szekvencia-elemet (bázist) pedig, amelyik tartalmazza a II. osztály valamelyik SA elemét, zöld háttérrel emeltem ki az I. osztály szekvenciák közül. A vastag, zöld színű SA elemek konszenzusa nem más, mint az I. osztály diszkrimináló elem (DE) készlete (A) ábra alsó sora, szintén vastagított zölddel kiemelve). Ha ilyenekből több is van, akkor a konszenzusnak megfelelő, egybetűs IUPAC kódot használtam. A B) ábrán ugyanez az elv látható, azzal a különbséggel, hogy ott a II. osztályba tartozó tDNS molekulái alatt pirossal szerepelnek az I. osztály SA elemei, illetve a szekvenciák között azok a bázisok, amelyek az I. osztály SA elemeivel megegyeznek az adott pozícióban piros háttérrel kiemeltek.

Sárga háttérrel a „*fals pozitív*" szekvenciákat emeltem ki. Ezeket egyik osztály DE elemét sem hordozzák[66].

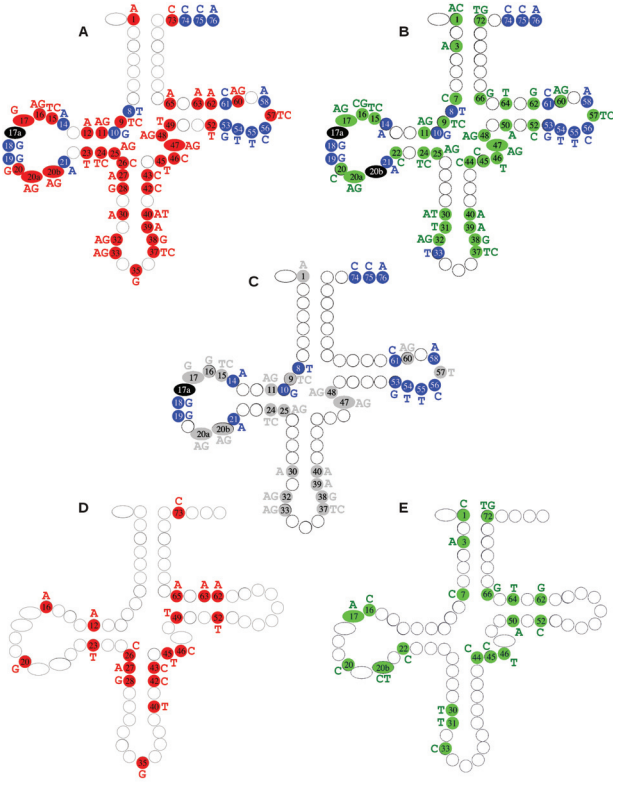

A
Élesztő I. osztály
B
Élesztő II. osztály
Cys GCA
Glu TTC
Glu TTC
Glu CTC
Ile TAT
Ile TAT
Ile AAT
Leu TAG
Leu GAG
Leu TAA
Leu CAA
Met CAT
Gln TTG
Gln TTG
Gln TTG
Gln CTG
Arg TCT
Arg CCT
Arg CCG
Arg ACG
Arg ACG
Val TAC
Val CAC
Val AAC
Val AAC
Trp CCA
Tyr GTA
II. osztály
ECP
„Strictly
absent"
Konzenzus
Ala TGC
Ala AGC
Asp GTC
Phe GAA
Phe GAA
Gly TCC
Gly GCC
Gly CCC
Gly CCC
His GTG
Lys TTT
Lys CTT
Asn GTT
Pro TGG
Pro TGG
Pro AGG
Ser GCT
Ser TGA
Ser CGA
Ser AGA
Thr TGT
Thr TGT
Thr CGT
Thr AGT
I. osztály
ECP
„Strictly
absent"
Konzenzus

láson az 5.2 ábrán, ahol az egyes osztályok SA illetve SP elemeit, valamint az egyes osztályokhoz tartozó DE-ket tüntettem fel. Tanulmányozható rajta ezek elhelyezkedése a két osztályban, amit ebben az ábrázolásban könnyen össze is lehet hasonlítani. Ezen az ábrán az élesztő tDNS szekvenciáiból származó adatok szerepelnek. A 5.2 ábrából megérthető, illetve vizualizálható az ECP logikája: a D) ábrát (II. osztály DE) úgy kapjuk meg, hogy az A)-n szereplő elemekből (I. osztály SA elemek) kivonjuk a C) elemeit (halmazelméletileg az I. osztály és II. osztály metszete). A másik osztálynál ugyanezt tesszük fordítva.

A 5.2 ábrát összehasonlíthatjuk a már korábban ismert és publikált eredményekkel [67]. A minden eukarióta tRNS-en megtalálható, közös elemeket (lásd még később) a mindkét osztályra jellemző SP elemek között figyelhetjük meg (C ábra).

5.1.2. Az SCP és ECP összehasonlítása

A fent leírt analízist a rövid szekvenciák illetve az élesztő tDNS szekvenciái után elvégeztem a tRNomics adatbázis 50 faján is illetve a *„Módszerek"* fejezetben leírtak szerint összehasonlítottam az ECP és SCP módszer hatékonyságát. Az analízisben 1210 I. osztályú és 1129 II. osztályú tDNS szekvencia vett részt.
Első megközelítésben a *„fals pozitív"* szekvenciák számát hasonlíthatjuk össze, amelyet a 5.1 táblázatban mutatok be. Mint említettem, minél kisebb ezek száma, annál relevánsabb lehet a két osztályba történt szétválasztás. Az SCP módszer, amely csak olyan pozíciók alapján osztályoz, amelyekben az adott osztály minden szekvenciája azonos elemet tartalmaz, az összes vizsgált szekvencia közül az I. osztály esetén 77%, a II. osztály esetén 88%-os fals pozitív arányt produkált. Az ECP esetében ez 17,5% illetve 18,5% volt, ami körülbelül ötödannyi, mint az SCP-nél. Ha az egyes fajokat nézzük, akkor a *„fals pozitív"* szekvenciák átlagos darabszáma az I. osztály esetén $4{,}2 \pm 2{,}2$, a II. osztály esetén $4{,}3 \pm 4{,}5$ míg ugyanezek az értékek az SCP-nél $20{,}9 \pm 10{,}0$ (I. osztály) és $17{,}7 \pm 10{,}4$ (II. osztály). Megjegyzendő azonban, hogy noha hat faj esetében az ECP tökéletesen definiálni tudta az adott osztályt (az adott osztályra nézve nem produkált *„fals pozitív"* szekvenciát), tökéletesen mégsem tudta a két osztályt egyetlen faj esetében sem szétválasztani, ugyanis nem találtunk olyan esetet, hogy mindkét osztály egyszerre lenne mentes a *„fals pozitív"* szekvenciáktól. Mindemellett – ahogyan a 5.1 táblázat is mutatja – az SCP módszer ezt az eredményt megközelíteni sem tudta.



5.2. ábra. Élesztőből származó adatokkal végzett ECP analízis eredménye a tRNS két dimenziós szerkezetén
Az A) ábrán az I. osztály, a B) ábrán a II. osztály SP elemeit kékkel, illetve az SA elemeket rendre pirossal és zölddel ábrázoltam. A C) ábrán az élesztő mindkét osztályára jellemző elemek szerepelnek, a közös SP elemek kékkel (azok az elemek, amelyeket mindkét osztály minden szekvenciája tartalmaz, és csak azt tartalmazza abban a pozícióban). A közös SA elemek szürkével jelöltek, ez azt jelenti, hogy a feltüntettet nukleotidokat egyetlen egy tDNS szekvencia sem tartalmazza az élesztőben. Az D) ábrán az I. osztály specifikus SA elemek (tehát azok az elemek, amelyek csak az I. osztályból hiányoznak) pirossal jelöltek. Ezek nem mások, mint a II. osztály diszkrimináló elemei. Az E) ábrán a II. osztály specifikus SA elemek zölddel jelöltek, ezek nem mások, mint a I. osztály diszkrimináló elemei. A feketével jelölt pozíciók hiányoznak az élesztő tRNS génjeiből, itt minden szekvenciában *„gap"* et találunk[66].

5.1. táblázat. A tDNS osztalyozás hatekonyságának matematikai analízise

	I. osztály					II. osztály				
		„fals pozitív" szekvenciák száma		Valószínűség (p)			„fals pozitív" szekvenciák száma		Valószínűség (p)	
	Szekvenciák száma	SCP	ECP	SCP	ECP	Szekvenciák száma	SCP	ECP	SCP	ECP
Saccharomyces cerevisiae	27	24	3	**0.17**	0.34	24	26	2	1.00	**0.21**
Schizosaccharomyces pombae	27	29	5	1.00	0.36	30	26	10	**0.11**	0.81
Caenorhabditis elegans	56	46	10	0.36	0.44	60	56	18	0.78	0.86
Drosophila melanogaster	44	31	4	**0.11**	0.81	34	44	8	1.00	0.89
Homo sapiens	60	57	34	0.89	**0.13**	58	43	12	**0.07**	0.55
Encephalitozoon cuniculi	22	22	2	0.86	**0.20**	23	22	8	0.61	0.91
Arabidopsis thaliana	75	63	1	0.60	**0.03**	71	54	1	0.38	**0.03**
Methanopyrus kandleri	18	8	2	**0.15**	**0.22**	15	8	3	**0.04**	**0.18**
Pyrococcus abyssi	25	20	2	0.58	0.26	20	16	2	0.39	**0.20**
Pyrobaculum aerophilum	23	21	3	0.91	**0.19**	22	15	6	0.44	0.53
Aeropyrum pernix	25	19	6	0.51	0.43	20	21	12	1.00	0.91
Archaeoglobus fulgidus	25	19	3	0.50	0.77	20	16	4	0.64	0.86
Halobacterium sp. NRC-1	25	16	2	**0.04**	0.31	20	25	3	1.00	0.26
Sulfolobus solfataricus	23	17	3	0.66	0.48	22	12	1	**0.23**	**0.17**
Sulfolobus tokodaii	23	20	3	0.89	0.31	22	16	3	0.46	0.28
Thermoplasma acidophilum	25	18	3	0.49	0.54	20	15	1	0.37	**0.13**
Ferroplasma acidarmanus	24	16	4	0.60	0.80	20	14	0	0.54	**0.05**
Methanosarcina barkeri	27	18	1	**0.04**	**0.13**	21	22	3	0.79	**0.23**
Methanococcus jannaschii	17	11	0	0.28	**0.20**	16	13	4	0.55	0.95
Methanobacterium thermoautotrophicum	20	13	2	0.44	0.66	16	14	3	0.77	0.68
Treponema pallidum	25	19	3	0.49	0.65	19	19	0	0.90	**0.02**
Borrelia burgdorferi	18	12	2	0.42	0.89	14	13	1	0.81	0.52
Chlamydia trachomatis	18	16	5	0.90	0.91	18	12	0	0.43	**0.10**
Synechocystis 6803	19	21	3	1.00	0.67	21	7	2	**0.06**	0.53
Anabaena	19	23	5	1.00	0.73	23	8	4	**0.05**	0.71
Lactococcus lactis	20	14	6	0.57	0.94	18	9	1	**0.09**	**0.20**
Listeria monocytogenes	19	13	1	0.41	0.29	20	15	6	0.72	0.96
Bacillus subtilis	23	16	4	0.55	0.63	21	17	2	0.76	**0.24**
Aquifex aeolicus	19	21	1	1.00	0.36	21	12	0	**0.18**	**0.15**
Mycobacterium tuberculosis	22	22	5	0.86	0.87	22	22	2	0.86	0.50
Deinococcus radiodurans	21	18	4	0.62	0.51	23	16	8	0.33	0.93
Neisseria meningitidis	22	20	7	0.74	0.97	20	14	6	0.37	0.89
Pseudomonas aeruginosa	20	21	5	1.00	0.61	21	13	5	0.27	0.67
Buchnera sp. APS	16	13	4	0.31	0.57	15	9	0	**0.03**	**0.17**
Bacillus halodurans	21	13	1	0.56	0.28	17	16	3	0.80	0.48
Thermotoga maritima	23	21	5	0.82	0.94	22	22	0	0.98	**0.19**
Campylobacter jejuni	19	12	4	0.28	0.89	15	12	1	0.34	**0.20**
Vibrio cholerae	25	22	2	0.58	0.42	22	17	3	0.35	0.46
Clostridium perfringens	20	18	3	0.55	0.65	18	20	1	1.00	**0.17**
Helicobacter pylori	19	13	2	0.56	0.66	16	11	1	0.32	0.32
Ralstonia solanacearum	20	23	6	1.00	0.91	23	13	2	**0.17**	0.46
Mycoplasma genitalium	18	17	6	0.89	0.98	17	14	1	0.62	**0.23**
Mycoplasma pneumoniae	19	17	5	0.89	0.94	17	14	1	0.61	**0.23**
Ureaplasma urealyticum	16	11	3	0.52	0.93	13	13	0	0.90	0.27
Xylella fastidiosa	22	22	5	0.98	0.64	22	15	2	0.46	**0.17**
Haemophilus influenzae	19	18	6	0.77	0.98	18	14	2	0.36	0.51
Escherichia coli	22	21	6	0.75	0.81	21	16	5	0.35	0.68
Rickettsia prowazekii	16	15	2	0.76	0.79	15	12	1	0.68	0.48
Yersinia pestis	21	22	7	1.00	0.86	22	15	4	0.32	0.61
Sinorhyzobium meliloti	43	22	22	1.00	0.60	22	12	24	**0.00**	0.39

5.1.3. Az ECP analízis osztályspecifikus diszkrimináló elemei

A fentiekből arra következtethetünk, hogy az ECP módszerrel előállított DE-k alkalmasak lehetnek arra, hogy a tDNS szekvenciák szintetáz osztályuk szerinti elválasztását a puszta konszenzus pozíciók figyelembevételénél jóval hatékonyabban megtegyék. Egy adott osztály DE-je egyértelműen elválasztja, „kizárja" azokat a (például egy másik osztályból vett, vagy akár ismeretlen eredetű) szekvenciákat, amelyek a megfelelő pozícióban tartalmazzák a „tiltott" elemet. A vizsgált 50 fajra jellemző eredményeket ahhoz, hogy vizuálisan megfelelően összehasonlíthatóak legyenek egyetlen táblázatba rendeztem úgy, hogy a két osztályra külön-külön, bontva soroltam föl csak a DE-ket, minden egyes vizsgált pozícióban (tehát a variábilis hurok itt sincsen megjelenítve) a 5.3 ábrán. A DE-k megjelenítéséhez itt is a IUPAC kódokat használtam.

Az eredmények értékelése során beszélhetünk faj-specifikus DE készletről , amelyeknek azokat az elemeket neveztem, amely egy adott fajra tartoznak és mindkét osztály DE készletét magába foglalják. Munkám során ezekre az elemekre fókuszáltam, mivel a klasszikus „konszenzus" elemeket (SP elemek) a kvázi „SCP módszer"-t alkalmazó korábbi in silico munkák [67] már jól feltárták.

A DE-k közül az alábbiakban elsősorban azokat emelem ki, amelyek olyan pozíciókban jelennek, meg, amelyek ismert identitáselemet hordoznak valamelyik modell fajban illetve azokat, amelyek abban az élőlénycsoportban, amelybe az adott faj tartozik, általánosan megjelennek: tehát az *E. coli* esetében a baktériumok, az élesztőnél az eukarióta fajok esetében. Az ősbaktériumok kevésbé ismert identitáselemei miatt erre a csoportra szűkebb értelemben vett megállapításokat nem teszek.

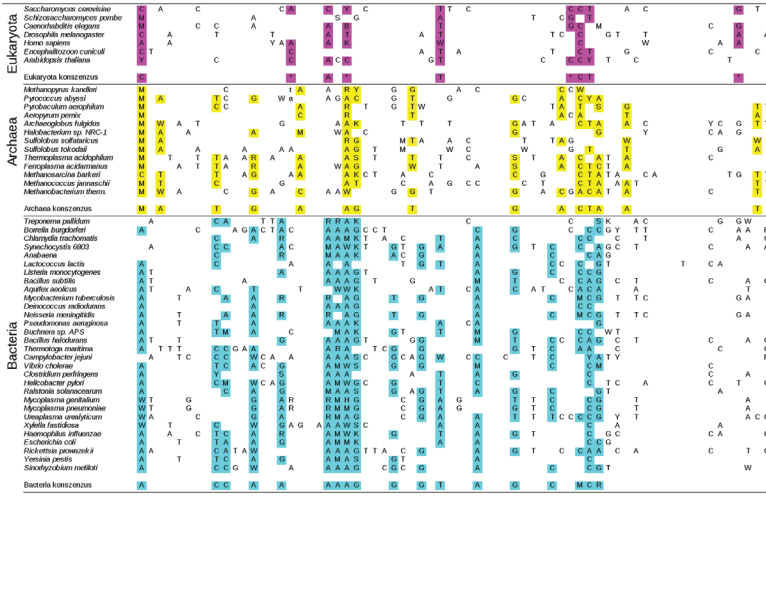
5.1.3.1. Az I. osztály diszkrimináló elemei

Két, a vizsgált fajokra általános DE-t hordoznak az I. osztályba tartozó tDNS szekvenciák. Az egyik az antikodon tripletjének középső bázisa, a 35-ös pozíció, amely a I. osztályban soha nem lehet G. Tehát az a szekvencia, amely G35-öt tartalmaz, az biztosan II. osztályba tartozik. Ilyenek természetes módon azok az aminosav-identitású tRNS molekulák, amelyek antikodon-tripletjének középső bázisa G, rendre a $tRNS^{Ser}$ (NGA),a $tRNS^{Ala}$ (NGC), a $tRNS^{Pro}$ (NGG) és

---

5.3. ábra *(lásd túloldalt).* Az ECP analízis diszkrimináló elemei
Az I. és a II. osztály DE-it a táblázat két külön része tartalmazza. A megjelentett pozíciók a Sprinzl-féle számozást követik, és nem tartalmazzák a variábilis hurkot és a o. pozíciót valamint a CCA véget sem. Színes háttérrel azok a trendek vannak kiemelve, amelyek az adott csoportra (magenta: eukarióta, sárga: ősbaktériumok, kék: baktériumok) jellemzőek, tehát a bennük szereplő fajok többségében megtalálható. Egy-egy DE, vagy a DE-k trendje nemcsak egy-egy nukleotidot, hanem nukleotidok csoportját is jelentheti, ennek megjelenítéséhez az IUPAC kódjait használtam.


I. osztály
Eukaryota
Archaea
Bacteria
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Caenorhabditis elegans
Drosophila melanogaster
Homo sapiens
Encephalitozoon cuniculi
Arabidopsis thaliana
Eukaryota konszenzus
Methanopyrus kandleri
Pyrococcus abyssi
Pyrobaculum aerophilum
Aeropyrum pernix
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Sulfolobus solfataricus
Sulfolobus tokodaii
Thermoplasma acidophilum
Ferroplasma acidarmanus
Methanosarcina barkeri
Methanococcus jannaschii
Methanobacterium therm.
Archaea konszenzus
Treponema pallidum
Borrelia burgdorferi
Chlamydia trachomatis
Synechocystis 6803
Anabaena
Lactococcus lactis
Listeria monocytogenes
Bacillus subtilis
Aquifex aeolicus
Mycobacterium tuberculosis
Deinococcus radiodurans
Neisseria meningitidis
Pseudomonas aeruginosa
Buchnera sp. APS
Bacillus halodurans
Thermotoga maritima
Campylobacter jejuni
Vibrio cholerae
Clostridium perfringens
Helicobacter pylori
Ralstonia solanacearum
Mycoplasma genitalium
Mycoplasma pneumoniae
Ureaplasma urealyticum
Xylella fastidiosa
Haemophilus influenzae
Escherichia coli
Rickettsia prowazekii
Yersinia pestis
Silonhyzobium meliloti
Bacteria konszenzus


II. osztály
Eukaryota
Archaea
Bacteria
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Caenorhabditis elegans
Drosophila melanogaster
Homo sapiens
Encephalitozoon cuniculi
Arabidopsis thaliana
Eukaryota konszenzus
Methanopyrus kandleri
Pyrococcus abyssi
Pyrobaculum aerophilum
Aeropyrum pernix
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Sulfolobus solfataricus
Sulfolobus tokodaii
Thermoplasma acidophilum
Ferroplasma acidarmanus
Methanosarcina barkeri
Methanococcus jannaschii
Methanobacterium therm.
Archaea konszenzus
Treponema pallidum
Borrelia burgdorferi
Chlamydia trachomatis
Synechocystis 6803
Anabaena
Lactococcus lactis
Listeria monocytogenes
Bacillus subtilis
Aquifex aeolicus
Mycobacterium tuberculosis
Deinococcus radiodurans
Neisseria meningitidis
Pseudomonas aeruginosa
Buchnera sp. APS
Bacillus halodurans
Thermotoga maritima
Campylobacter jejuni
Vibrio cholerae
Clostridium perfringens
Helicobacter pylori
Ralstonia solanacearum
Mycoplasma genitalium
Mycoplasma pneumoniae
Ureaplasma urealyticum
Xylella fastidiosa
Haemophilus influenzae
Escherichia coli
Rickettsia prowazekii
Yersinia pestis
Silonhyzobium meliloti
Bacteria konszenzus

a $tRNS^{Thr}$ (NGT antikodonokkal). Ezek aaRS-i az ismert módon a II. osztályba tartoznak.
A másik, 50-ből 47 fajra jellemző DE a C73, amely a 73-as, úgynevezett „diszkriminátor" pozícióban van [72]. Ez fontos identitáselem, a legfontosabb, legtöbbször identitáselemet hordozó pozíció az antikodon három pozíciója mellett. A bakteriális és ősbakteriális adatcsoportban a C73 a $tRNS^{His}$ identitáseleme, illetve az eukarióta csoportban a tRNSPro-ra jellemző. Mindkét identitás a II. osztályba tatozik.
A bakteriális adatcsoportban jelenik meg a DE-k között az C1:G72 bázispár, ez a $tRNS^{Pro}$-ban ismert identitáselem, amely a II. osztályba tartozik. Fontos megjegyezni, hogy ezt az I. osztály specifikus DE-t nem találjuk meg az eukarióta csoportban, ami nem lehet meglepő, hiszen az élesztőnél leírták ezt a bázispárt, mint identitáselem a $tRNS^{Tyr}$-nál, amely viszont az I. osztályba tartozik.

5.1.3.2. A II. osztály diszkrimináló elemei

A II. osztályban nem figyelhetünk meg kiemelt, identitáselemet hordozó pozíción DE-t. A 45-ös pozícióban, ahol a C45 DE, a tRNSPhe esetén (amely a II. osztályba tartozik) a T45 identitáselem. Ezzel tehát nem szül ellentmondást: a II. osztály akár kizárhatja a C45-öt. Emellett a T46 illetve tágabb értelemben az Y46 szinte minden fajban megjelenik, mint DE. Érdekesebb megfigyeléseket tehetünk, hogyha az élővilág egyes *„kingdom"*-jait vesszük sorra.
A bakteriális szekvenciák közt általános ebben az osztályban az A1:T72 DE pár. Ezek *E. coli* $tRNS^{Trp}$-ben és $tRNS^{Gln}$-ben identitáselemek, amelyeket élesztőben nem írtak le. Mindkét identitás az I. osztályba tartozik. Ezzel szemben az 1G:C72, amely ebben a pozícióban megengedett, a II. osztályba tartozó tRNS-ek, a Thr és a Gly identitáselemei, előbbi az élesztőben is. Megfigyelhetjük tehát, hogy ebben a pozíciópárban az élesztő és a 5.3 ábrán látható módon az eukarióta és ősbakteriális fajok többsége eltérő identitás- és DE-készletet használ. Az ősbaktériumok egyértelműen a M1:K72 párt, az eukarióta fajok ehhez kicsit hasonlóan, az 1. pozícióban főleg C illetve A (vagy mindkettőt: M) a 72-ben K-t (az M párjait: G illetve T) zárnak ki. Tehát ez utóbbi két élőlénycsoport a legtöbb faj esetében a baktériumoknál „szigorúbb" szabályokat használ.
Identitáselemet hordozó pozíciókban csak a bakteriális szekvenciáknál fordul elő DE, az A34, azonban az A34-et, mint identitáselemet még nem írták le.

5.1.3.3. Rejtett, potenciális osztályspecifikus elemek

A 5.3 ábrán a fent említetteken túl jó néhány, számos fajban megtalálható, eddig nem említett DE-t láthatunk. Ezek többsége olyan pozícióra esik, amelyek eddigi ismereteink szerint nem

hordoznak identitáselemeket egyik osztályban sem. Ilyen DE-ket sok esetben ún. „opcionális" pozícióban, változó hosszúságú (gyakran *„gap"*-es) régiókban, pl. a D-karon (17, 17a, 20, 20a, 20b pozíciók), illetve a variábilis hurok környékén (44-47 pozíciók) találtam. Noha ezek a pozíciók nem hordoznak ismert identitáselemeket, a tRNS szerkezetének kialakításában azonban jelentőségük lehet. A szintetázzal közvetlen kapcsolatba nem kerülő, ezért az irodalomban *„cryptic"*, rejtett elemeknek nevezett [24] elemeket feltételezhetünk ezekben a pozíciókban. Ezek az elemek az identitás kialakításában közvetlenül nem vesznek részt, ugyanakkor fontos, osztályspecifikus jelleggel is bírhatnak.

5.1.3.4. Közös, osztályspecifikus DE elemek az élővilág nagy csoportjaiban

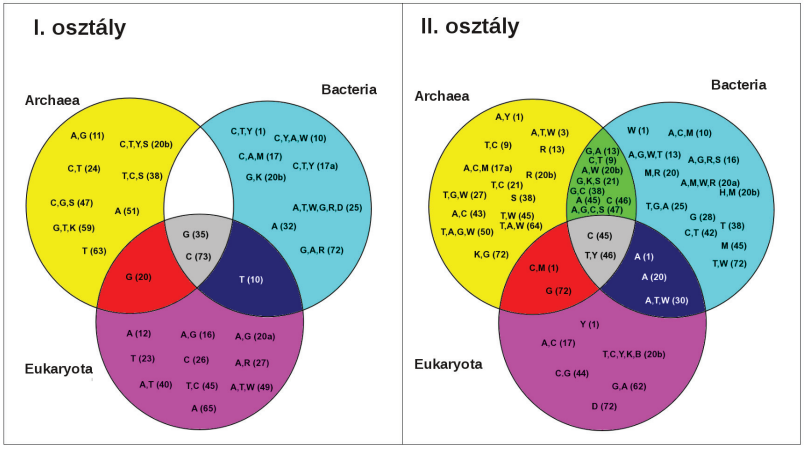
Ahogyan azt már több példán bemutattuk, a DE elemeken különböző mértékben „osztoznak" a vizsgálatban részt vett fajok, illetve fajcsoportok. Ahhoz, hogy a közös tulajdonságokat, amelyeket a 5.3 ábra is kiemel, még szemléletesebben tudjam bemutatni, Szathmáry Eörs javaslatára Venn-diagramon ábrázoltam a 5.4 ábrán. Itt jól megfigyelhetőek, hogy melyek azok a DE-k, amelyeket csak a baktériumok, csak az ősbaktériumok, vagy csak az eukarióta fajok használnak, és melyek azok a pozíciók, amelyek az élővilág egyes *„kingdom"*-párjaira jellemző DE-ket hordoznak.

5.1.4. Az ECP osztályokat szétválasztó képessége

A *„Módszerek"* fejezetben írtak szerint megvizsgálhatjuk, hogy az ismertetett DE-k statisztikailag mennyire relevánsak, illetve azt, hogy az ECP módszer a különböző megadott szempontok szerint (a már ismertetett, *„fals pozitív"* szekvenciák darabszámán túl) mennyiben ad más eredményeket, mint az SCP.

A 5.1 táblázatban foglaltam össze az SCP és ECP analízisek eredményeit. A *„fals pozitív"* szekvenciák darabszáma mellett azok arányát is figyelembevettem, és a határt (szignifikancia szintet) <= 25%-nál húztam meg. Az SCP a vizsgált 100 tDNS szekvencia-csoportból (50 faj, két osztály) mindössze 16 esetben (5 I. és 11 II. osztályba tartozó csoportnál) adott szignifikáns eredményt, míg az ECP ennél 60%-kal jobban teljesített, 27 esetben (7 I. és 20 II. osztályba tartozó csoportnál). Az ECP szelektivitása azonban a vizsgált fajcsoportok közül eltérő a két osztályban: az eukarióta és ősbaktérium fajoknál a két osztályban megegyezik, a baktériumoknál viszont az I. osztály esetén egyszer sem adott szignifikáns értéket.

Ez a viszonylag alacsony hatékonyság azt jelezheti, hogy az egyes aminosav-identitások nagymértékben szét vannak szórva a szekvenciatérben. Emiatt lehet az, hogy eredeti (a priori), a szintetázuknak megfelelően képzett két csoport ECP DE-k alapján képzett szeparáltsága alig tér



5.4. ábra. Az ECP analízis diszkrimináló elemei az élővilág három nagy doménje szerint bontva

Az I. és a II. osztály DE-it, a trendszerűen megjelenő elemeket csoportosítva mutatja be az ábra, attól függően, hogy az élővilág nagy csoportjai (színezésben is megegyezően a 5.3 ábrával: magenta: eukarióta, sárga: ősbaktériumok, kék: baktériumok) közül melyeket használ egyik vagy másik csoport közösen. Ilyen formán a teljes élővilágban megegyező diszkrimináló elemek a szürke hátterű, közös metszetben szerepelnek.

el a véletlenszerűen létrehozott csoportokétól.

5.1.5. Egyedi, osztályspecifikus DE-készletek

Ha a DE elemek nem, vagy csak kis mértékben definiálják az a priori osztályokat, akkor feltehetjük a kérdést: vajon ezek az elemek mennyire jellemzőek az adott osztályra? Ezt véletlenszerűen létrehozott osztályok segítségével mondhatjuk meg. Az összes lehetséges létrejövő osztályból megvizsgálhatjuk, hogy hány esetben keletkezik ugyanaz DE-készlet, mint az a priori esetében.
50 fajból 29 esetben az ECP által létrehozott DE-k egyszer sem ismétlődtek meg. A maradék 21 fajból 16-ban négy, vagy annál kevesebb véletlenszerűen létrehozott két csoport eredményezett csak ugyanolyan DE készletet, mint az a priori osztály. Az ECP ennél kevésbé volt hatékony *Neisseria meningitidis* esetén, ahol 55-ször, az *Aeropyrum pernix* esetén 34-szer, a *Pseudomonas aeruginosa* esetén 19-szer, a *Deinococcus radiodurans* esetén 14-szer és a *Yersinia pestis* esetében 9-szer eredményezett ugyanolyan DE-készletet az analízis.
Ugyanakkor az irodalomból ismerjük, hogy az SCP módszer erre nem képes. Ugyanezt az analízist elvégezve a legjobb eredmény a Methanopyrus kandleri estében kapjuk, ahol 125 véletlenszerű esetben kapjuk „csak" ugyanazokat az SP elemeket.
Ebből is látszik, hogy az ECP kiemelten alkalmas arra, hogy osztályspecifikus nukleotidokat, nukleotid csoportokat (ún. DE-ket) találjunk segítségével.

5.1.6. Egyedi DE-k

Miután beláttuk, hogy a DE-készletek az a priori osztályokra jellemzőek, megvizsgálhatjuk, hogy az egyes DE-k külön-külön mennyire jellemzőek az adott osztályra. A *„Módszerek"* fejezetben leírtak szerinti analízis eredményét a 5.2 táblázatban mutatom be.

A táblázatból jól látszik az I. osztályban leírt G35 DE, illetve a II. osztály T72 (tRNS-ben U72) is több fajban előfordul. A legtöbb felsorolt elem inkább egy-egy fajra jellemző, ezek közül – az ismert modellfajokra fókuszálva – kiemelendő az élesztő I. osztálynál az A12, amely a II. osztályhoz tartozó szintetázú $tRNS^{His}$-re jellemző. Hét vizsgált fajban is megjelenik a II. osztály C34-es pozíciója, amely valószínűleg csak ezekre a fajokra lehet jellemző (a legtöbb baktérium például itt A34-et használ DE-ként).

5.2. táblázat. Az osztályok jellemző SA („*strictly absent*”) elemei

Faj	I. osztály	II. osztály
Saccharomyces cerevisiae	G35	
Schizosaccharomyces pombae	A6, G35, U67	
Caenorhabditis elegans	G35	
Drosophila melanogaster	G35	
Homo sapiens	G35, A52	
Encephalitozoon cuniculi	G35, G44	A71, U2
Arabidopsis thaliana	G28, G35, G50, C42	G32, C41
Methanopyrus kandleri	G35, U32	
Pyrococcus abyssi	G35	G31, C39
Pyrobaculum aerophilum	G35	A17a
Aeropyrum pernix	G35	
Archaeoglobus fulgidus	G35	
Halobacterium sp. NRC-1	G35	C17a
Sulfolobus solfataricus	G35	
Sulfolobus tokodaii	A42, G35, U20a, U28	
Thermoplasma acidophilum	G35	A43, U27
Ferroplasma acidarmanus	G35	A17a, A27, A43, U20b, U27
Methanosarcina barkeri	G35	U65
Methanococcus jannaschii	G35	C34
Methanobacterium thermoautotrophicum	G35	
Treponema pallidum	G35	A51, A63, U63
Borrelia burgdorferi	G35	C34
Chlamydia trachomatis	G35	
Synechocystis 6803	G35	
Anabaena	G35	
Lactococcus lactis	G35	
Listeria monocytogenes	A6, G35, U67	
Bacillus subtilis	G35	G27, C34, C43
Aquifex aeolicus	G35, U65	A51, U63
Mycobacterium tuberculosis	G35	
Deinococcus radiodurans	G35	
Neisseria meningitidis	G35	A24, U11
Pseudomonas aeruginosa	G35	U72
Buchnera sp. APS	G35	C34, U72
Bacillus halodurans	G35	C34
Thermotoga maritima	G35	
Campylobacter jejuni	G35	A13, A27, C34
Vibrio cholerae	G35, U59	C34, U72
Clostridium perfringens	G35, U45	A27, C16
Helicobacter pylori	A42, G35	A13, C34
Ralstonia solanacearum	G35	
Mycoplasma genitalium	G35	G6, C67
Mycoplasma pneumoniae	G35	G6, C67, U40
Ureaplasma urealyticum	G35	C46, C47, U45
Xylella fastidiosa	G35	U3
Haemophilus influenzae	G35, U59	
Escherichia coli	G35	A13
Rickettsia prowazekii	G35	A51
Yersinia pestis	G35	
Sinorhyzobium meliloti	A50, G35, C17	

5.1.7. Az ECP módszer értékelése

Korábban csak olyan módszereket alkalmaztak, amelyek csak az identitások vagy identitás csoportok (például szintetáz osztályok) szigorúan, minden szekvenciában jelenlévő nukleotidjain alapultak. Az ECP módszer kiterjesztette a korábbi megközelítést azáltal, hogy azokat a nukleotidokat is figyelembeveszi (sőt ezekre fókuszál), amelyek adott csoportokból hiányoznak. A „szigorúságot" emellett megtartotta, ami azt jelenti, hogy minden egyes szekvencia számít ebben az osztályozásban. Ebből adódóan a módszer érzékeny, tehát minden egyes szekvencia „beleszól" végső eredménybe. Ez a tRNS-ek és tRNS gének esetében indokolt, hiszen – főleg a baktériumok és ősbaktériumok esetében – kevés számú tRNS izoakceptorral dolgozik egy-egy szervezet.

A statisztikai eredményekből láthattuk, hogy az ECP módszer a korábbi SCP megközelítésnél hatékonyabb, alkalmas osztályspecifikus elemek feltárására. Megjegyzendő azonban, hogy itt (és a további elemzésekben) nem vesz figyelembe poszt-transzkripciós módosításokat.

A legfontosabb eredmény az, hogy a korábbi nézettel szemben vannak olyan elemek, amelyek jellemzőek az egyik, illetve másik tRNS osztályra, bizonyítva ezzel azt a feltételezést, hogy a tRNS-ek a nekik megfelelő szintetázzal koevolválódtak. Ennek bizonyítékai eddig azért maradhattak feltáratlanok, mert a keresések csak az egyes pozíciókban meglévő nukleotidokra irányultak, és nem vették figyelembe azt, hogy vannak-e törvényszerűségek a hiányzó nukleotidok előfordulásában. Vizsgálatunkkal arra a következtetésre jutottunk, hogy bizonyos pozíciók egy-egy tRNS osztályra jellemzően valamelyik nukleotidot vagy nukleotid csoportokat nem engednek meg. Ez a tRNS szerkezetének kialakítását, a tRNS-szintetáz kapcsolatot, végső soron a tRNS identitását befolyásolják kisebb vagy nagyobb mértékben. Ennél fogva az ECP módszer alkalmas lehet arra, hogy a DE-kel olyan pozíciókat, nukleotidokat tárjunk fel, amelyek megakadályozzák, hogy egy oda nem illő szintetáz tévesen töltse föl az adott tRNS-t egy rossz aminosavval: tehát alkalmas lehet a módszer antideterminánsok helyének predikciójára.

5.1.8. Az osztályspecifikus elemek kísérleti eredmények tükrében

Az irodalomban ismeretesek olyan mutációs kísérletek, amelyek segítségével úgy tárják föl a tRNS identitáselemeit, hogy egy-egy pozícióban ismert vagy feltételezett identitáselemeket (nukleotidokat) cserélnek ki egy másik tRNS identitáselemeire. Ezeket akár in vivo vagy in vitro rendszerben a cél tRNS molekula szintetáza tölti fel az új aminosavval. Ezek az ún identitásváltó („*identity switch*") kísérletek. A mi szempontunkból érdekes kísérlet az a speciális eset, amikor az identitás-váltás két különböző osztályba tartozó tRNS molekula között történik. Ilyenre az irodalomban szinte alig akad példa. McClain [73] és munkatársai közöltek

egy ilyen kísérletet a tRNS esetében. A II. osztályú $tRNS^{Gly}$ identitáselemei az U73, G1:C72, C2:G7, G3:C70 és C35, amelyeket más „fogadó" tRNS-ekbe illesztettek. Ezek a $tRNS^{Phe}$ és $tRNS^{Lys}$ voltak a II. illetve a $tRNS^{Arg}$ és $tRNS^{Gln}$ az I. osztályból. Az utóbbi kettő identitásváltó kísérletben tehát az osztály is változott. A kísérletek eredményeként – az összes identitáselemet átültetve – a kiindulási tRNS-ek Gly-nel töltődtek föl. Bármelyik elemet elhagyva nem sikerült teljes identitásváltást elérni.

Analízisünk eredményeiből látszik, hogy ezek az identitás pozíciók nem tartalmaznak osztályokat definiáló DE-t. Kérdés, hogy ez nem vezet-e ellentmondáshoz. Át lehet lépni az osztályhatárt úgy, hogy közben nem érintünk osztályspecifikus DE-ket? Az I. osztályba tartozó tRNS esetében azok között a pozíciók között, amelyeket a kísérlet nem érintett, tehát ahol az eredeti nukleotidok maradtak meg, ott vannak azok is, amelyek osztályspecifikus DE-ket tartalmaznak. Ezeknek a II. osztályba tartozó Gly-aaRS-t „távol kellene tartaniuk".

Ennek a gondolatnak a nyomán megvizsgáltam, hogy milyen arányban vannak jelen a II. osztályba tartozó $tRNS^{Gly}$ identitáselemek az I. illetve II. osztályba tartozó szekvenciákban.

Megvizsgálva az a 22 I. osztályba és a 18 II. osztályba tartozó *E. coli* szekvenciát (kihagyva a $tRNS^{Gly}$-ket) a következőket kaptam. Az U73 1 db I. osztályba tartozó szekvenciában van meg, a II. osztályból hiányzik; a G1:C72 18 db I. osztályú és 14 db II. osztályú szekvenciában; a C2:G7 9 db I. osztályú és 6 db II. osztályú szekvenciában; a G3:C70 11 db I. osztályú és 4 db II. osztályba tartozó szekvenciában van, míg a C35 6 db I. osztályú és 1 db II. osztályú szekvenciában szerepel. (Természetesen egyetlen I. osztályba tartozó nem-Gly identitású tRNS szekvencia sem tartalmazza egyszerre az összes $tRNS^{Gly}$ identitáselemet.) Abból, hogy a II. osztályba tartozó $tRNS^{Gly}$ identitáselemei gyakrabban fordulnak elő I. osztályú szekvenciákban mint II.-ban látszik, hogy az osztályspecifikus jellegeket kimutató ECP analízisünk nem képes azonosítani identitás specifikus elemeket. Más oldalról megközelítve a kérdést ugyanakkor elmondható, hogy az identitást meghatározó elemeknek természetesen nem csak az a feladatuk, hogy a másik osztályba tartozó tRNS-ektől megkülönböztessék az adott identitást. Ugyanezt meg kell tenniük a saját osztályukba tartozókkal szemben is. Az osztályokon végrehajtott ECP elemzésekből levezetett DE-készletek tehát nem az egyes identitásokat különböztetik meg egymástól, hanem a két különböző, szintetáz osztályhoz tartozó csoportokat.

5.2. Osztályspecifikus elemek feltárása *„logo”* módszerrel

Az eddig ismertetett eredmények tükrében arra voltam kíváncsi, hogy vajon osztályspecifikus elemeket (potenciális determinánsokat vagy antideterminánsokat) egy nem diszkrét módszerrel, a tDNS-ek *„function”* illetve *„inverse function logo”*-ival fel lehet-e tárni.
Ezekről az analóg módszerekről bebizonyosodott, hogy képesek bizonyos esetekben determinánsok feltárására, illetve a kevés ismert esetet figyelembe véve akár antideterminánsok felderítéséhez is [63]. Ahhoz, hogy az analízis az ECP módszerrel összevethető legyen, az antideterminánsok meghatározásához szükséges *„inverse function logo”*-kat kell képeznünk, hiszen a DE-k potenciálisan antideterminánsok lehetnek, funkciójuk az oda nem illő szintetázok távol tartása.

5.2.1. Az I. és a II. osztály *„inverse function logo”*-i

Ha az I. és a II. osztályba tartozó tDNS szekvenciákat különválasztjuk és a *„Módszerek”* fejezetben leírtaknak megfelelően egy *„profil mátrix”*-ot állítunk elő, akkor az I. illetve a II. osztálynak meg tudjuk rajzolni a *„function logo”*-it. Tehát itt az I. és a II. osztályt kvázi egy-egy identitásnak tekinthetjük. Így a szekvenciatér olyan módon alakul, mintha (a 20 aminosav helyett) összesen két identitás szerepelne benne. Ennél fogva az alábbi matematikai megállapításokat tettem: Mivel a két osztály mérete – bennük található szekvenciák száma – közel azonos, ezért a $p(1) = p(2)$, tehát az $I(1|x) = I(2|x)$. Az *„inverse function logo”* a magasságok így az infromáció tartalom az alábbi szerint alakul(vesd össze a 1.2 egyenlettel):

$$I_l(y|x) = 1 \tag{5.1}$$

A *„function logo”* magasságok tehát a 1.1 egyenlet alapján:

$$h_l(1|x) = p \tag{5.2}$$

$$h_l(2|x) = 1 - p \tag{5.3}$$

Az *„inverse function logo”* pedig a két osztályra az ún. reciprok inverz esetén [64]:

$$h'_l(1|x) = \frac{\frac{1}{p}}{\frac{1}{p} + \frac{1}{1-p}} = 1 - p \tag{5.4}$$

$$h'_l(2|x) = \frac{\frac{1}{1-p}}{\frac{1}{p} + \frac{1}{1-p}} = p \tag{5.5}$$

A gyakorlatban ez azt jelenti, hogy két osztály esetén az egyik osztály „*inverz logo*”-ja a másik „*function logo*”-ja. Tehát $|Y| = 2$ a két osztály esetében ($|Y| = \{1,2\}$) ha az „*inverz logo*”-k magassága h' és a „*function logo*”-k magassága h, akkor az alábbi összefüggés adódik:

$$h_l(1|x) = h'_l(2|x) \text{ illetve } h_l(2|x) = h'_l(1|x) \tag{5.6}$$

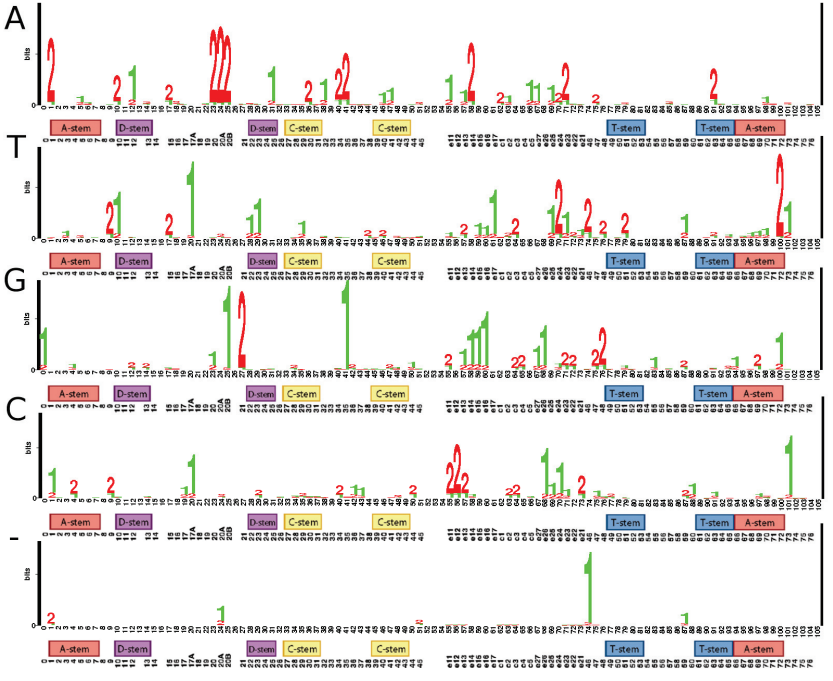
A fönti praktikus megfigyelés szem előtt tartva mutatom be a 5.5 ábrán az I. és a II. osztály bakteriális tDNS-einek „*inverse function logo*”-ját.

5.2.2. Az „*inverse function logo*”-k és a diszkrimináló elemek összefüggései

Mivel a „*logo*”-kat az MSDB adatbázisra készítettem el, összehasonlíthatjuk a 5.5 és a 5.3 ábra kékkel kiemelt (bakteriális szekvenciák DE-i) elemeit, figyelembe véve a már publikált, ugyanebből az adatbázisból készített, tRNS „*function logo*”-kal, amelyet a funlogoamino.png ábrán már bemutattam.
A legmarkánsabb, nem csak a bakteriális szekvenciákra jellemző DE-ket az „*inverse logo*” is jól láthatóan visszaadja: az I. osztály DE-i G35 és C73 egyértelműen látható „*logo*”-kat adnak. A bakteriális II. osztályra jellemző DE az A1:T72 bázispár szintén megjelenik a „*logo*”-kban. A 5.5 ábrán szintén jól kiemelkednek azok az opcionális pozíciók, amelyek az I. osztályra (Y17a, G20b) illetve a II. osztályra (A20, A20a, A20b) jellemző DE-k. Ez utóbbi példák közül a 17a pozíciónál meg kell jegyezni, hogy az I. osztályban itt a T és a C, a pirimidinek (Y) a tiltottak a bakteriális szekvenciák között. Ugyanezt az analízist *E. coli*-ra elvégezve csak T17a-t kapunk (lásd 5.3 ábra megfelelő sora), a többi baktériumfajjal kiegészítve az adatokat azonban a T mellett jelentős számban C is a tiltott elemek között szerepel.

5.2.3. Az I. és a II. osztály „*logo*”-inak értékelése

Ha a 5.5 ábrát és a funlogoamino.png ábrán egyes pozíciókat összehasonlítunk, megállapíthatjuk, hogy az egyes osztály-”logo”-k az adott osztályhoz tartozó aminosav-identitásokból „épülnek fel”. Tehát ha az egyes, azonos osztályba tartozó aminosavak magasságát összeadjuk, jó közelítéssel az osztály-”*logo*” méretéhez jutunk (természetes a „*logo*” számítás sajátosságai miatt ez az egyszerű logika nem minden esetben érvényesül): példaként kiemelve a C73 DE-nél az I. osztály „*inverse function logo*”-ja van, amely a II. osztályú $tRNS^{His}$ „*function logo*”-ja, vagy a G35, amelyik szintén az I. osztály DE-je, és a $tRNS^{Thr}$, $tRNS^{Pro}$, $tRNS^{Ala}$ és $tRNS^{Ser}$ „*function logo*”-ja, amelyek rendre mind a II. osztályba tartoznak. Ellenkező eset figyelhető meg az 1. pozíció esetében, ahol U1 „*logo*”-k figyelhetőek meg a $tRNS^{Gln}$ és $tRNS^{Asn}$ esetében, amelyek



5.5. ábra. Az I. és a II. osztály bakteriális szekvenciáinak *„inverse function logo”*-ja 1-essel (zöld színnel jelölve) azok a pozíciók jelennek meg, ahol a megfelelő bázis nem jellemző, hiányzik a I. osztályba tartozó bakteriális tDNS szekvenciákból. Ugyanígy 2-es számmal (piros színnel jelölve) azok a pozíciók jelennek meg, ahol a megfelelő bázis nem jellemző, hiányzik a II. osztályba tartozó bakteriális tDNS szekvenciákból.

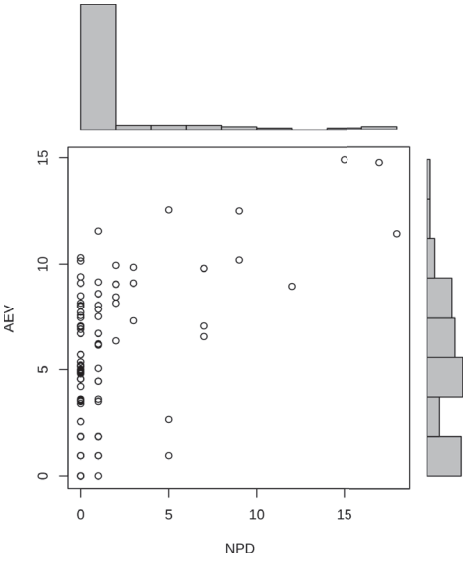
azonban két különböző osztályba tartoznak, ezért itt osztály-„*logo*”-t nem kapunk.
A mi diszkrét módszerünket összehasonlítva az analóg módszerrel kapott eredményekkel azt láthatjuk, hogy az sok esetben hasonló eredményekre jut. A diszkrét módszer előnye azonban az, hogy egyrészt gyorsabb, másrészt minden egyes szekvenciát figyelembe tud venni. Említettem, hogy ennek kiemelkedő biológiai relevanciája van, hiszen, ha azt felételezzük, hogy a szekvenciáink közül minden egyes tDNS működőképes, az élő szervezetben ténylegesen jelenlevő, a szintetáza által jól felismerhető tRNS molekulát eredményez, akkor olyan megállapításokat, törvényszerűségeket kell levonnunk (akár identitáselemekre vonatkozóan is), amelyek minden egyes vizsgált tDNS-re, így az alulreprezentáltakra is igazak lesznek. Itt mutatkozik meg a diszkrét módszer előnye. Ahhoz azonban, hogy a „*logo*”-khoz hasonló felbontású – aminosav-specifitás szintű – eredményeket állíthassak elő az ECP módszer segítségével, az eddigi eljárást tovább kellett fejleszteni. Mindemellett azt is el kellett érnem, hogy az ECP módszer érzékenysége mellett releváns eredményeket állítsak elő oly módon, hogy a „*logo*”-knál egyértelműbb jeleket kapjak. Ez utóbbi módszer hátránya ugyanis az, hogy elsősorban vizualizációs célokat szolgál.

5.3. Új identitás helyek feltérképezése tRNS pozíciók átlagos DE számának segítségével

A korábbi fejezetekben az ECP módszert felhasználva a két aminoacil-tRNS szintetáz osztálynak megfelelően elkülönített tDNS csoport (elsődleges) szekvenciális alapon történő elválasztását mutattam be. Az ECP-módszer „*strict*” logikáját felhasználva a következőekben a tDNS-eket aminosavspecifitásuk alapján különítem el és az elkülönítés során olyan pozíciókat határozok meg, amelyek a tRNS-ek identitásában szerepet játszhatnak egy vagy több specifitás esetében.
Ehhez a „*Módszerfejlesztés*” fejezetben bemutatott átlagos kizárási érték (AEV) fogalmát vezettem be és használtam fel. Ehhez először megállapítottam az egyes identitáspárokhoz tartozó DE elemeket a tRNAdbC adatbázis három alcsoportjára: a bakteriális, az eukarióta és az ősbakteriális adathalmazokra. Minden egyes pozícióra összegezve az AEV értékeit is kiszámítottam, amelyet a 2. számú melléklet mutat be, szintén mindhárom csoportra.

5.3.1. Az AEV statisztikai értékelése

Egyszerű statisztikai elemzést végeztem azért, hogy megállapítsam, vajon az egyes pozíciókban kapott AEV érték hogyan viszonyul a kísérletesen megállapított, eddig ismert identitáselemek

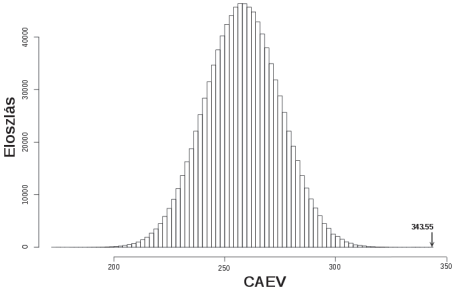


5.6. ábra. Az AEV értékek korrelációja az ismert identitáselemek számával
Az ábrán egy-egy kör, egy-egy pozíciót jelöl. A tengelyekről leolvasható, hogy egy bizonyos AEV-jű pozícióhoz mekkora NPD érték tartozik. Az egyes értékek eloszlását a tengelyekre vetített oszlopdiagramok mutatják.

számához (továbbiakban: NPD - *„number of published determinants”*). Ehhez minden egyes pozícióban számba vettem az eddig közölt identitás-elemeket, egészen pontosan csak a determinánsokat illetve a kiszámított AEV értékeket (lásd a pontos értéküket a 2. számú mellékletben). A statisztikai elemzést ugyanakkor csak az *E. coli* (Bacteria) szekvenciákon végeztem el, mivel kísérletes adatok csak itt állnak ehhez elégséges számban rendelkezésre, de a feltételezések szerint még vannak fel nem tárt identitáselemek, tehát az eddig kísérletesen megállapított identitáselemek a teljes identitáselem-készletnek csak egy részét jelentik. A két adatsoron korrelációs analízis végeztem el. A Pearson-korreláció mellett, amellyel a két adatsor linearitása igazolható. A Spearman-féle rangkorrelációt kiszámítása is indokolt, mert az NPD értékek eloszlásáról nem állíthatjuk, hogy normál eloszlású, az eloszlás erősen ferde és sok kiugró értékeket tartalmaz, amelyek a korrelációt erősen befolyásolják.

Az AEV és NPD adatsor korrelációját a 5.6 ábrán szemléltetem.

Számításaim alapján a kísérleti adatokból származó determinánsok és az elemzésünkből származó AEV értékek közepesen erős korrelációt mutatnak. Az adatsorok között csak



5.7. ábra. Az AEV értékek korrelációja az ismert identitáselemek számával
Az ábrán egy-egy kör, egy-egy pozíciót jelöl. A tengelyekről leolvasható, hogy egy bizonyos AEV-jű pozícióhoz mekkora NPD érték tartozik. Az egyes értékek eloszlását a tengelyekre vetített oszlopdiagramok mutatják.

kismértékű linearitást tapasztalható, a Pearson-korrelációjuk közepes (R=0,54). A trendek megállapítása Spearman-korrelációval történt, ahol szintén közepes összefüggést kaptam (ρ=0,54). A statisztikai elemzést elvégeztem egy szelektált adatsoron is. Itt csak azokat a pozíciókat vettem figyelembe, ahol legalább egy, már kísérletesen megállapított identitáselemet leírtak. Itt lényegesen erősebb összefüggést kaptam (R=0,67; ρ=0,60).

Mivel korrelációs analízis a teljesen random (R/ρ=0) és a tökéletes korreláció (R/ρ=1) közötti értéknek adódott, az eredmények értelmezéséhez egy bootstrap-jellegű analízist is elvégeztem. Megállapítottam, hogy azoknak a pozícióknak az AEV összege, amelyek tartalmaznak identitáselemet 343,55 (ez a CAEV - *„cumulative AEV"* érték). Ez után az összes pozícióra (96) számított AEV értékekből véletlenszerűen pontosan annyit vettem ki, amennyi azoknak a pozícióknak a száma, amelyek tartalmaznak ismert identitás elemet (40 darabot). Ezeket a számokat összeadtam. A továbbiakban ezt elvégeztem százezerszer, és megnéztem, hogy a kiválasztott 40 AEV érték összege milyen eloszlást ad (lásd 5.7 ábra). Utána megnéztem, hogy a 343,55 vagy ennél nagyobb összeg-értéknek mekkora a gyakorisága. Nem kaptam egyetlen ilyen esetet sem, ami jelezte, hogy az esemény valószínűsége 10-5-nél kisebb lehet. Az összes számba vehető eset (96 pozícióból 40 kiválasztása) 1027 nagyságrendű, míg a 340 feletti összegű eredmények számossága 102 nagyságrendű, a 340 fölötti összegek valószínűsége nagyságrendileg 10-24, tehát elhanyagolható.

Ezek az eredmények alátámasztják, hogy a magas AEV értékkel rendelkező pozíciók az

esetek legtöbbjében identitáselemet hordozó (magas NPD-jű) pozíciókra esnek.

5.3.2. Az AEV eredményei

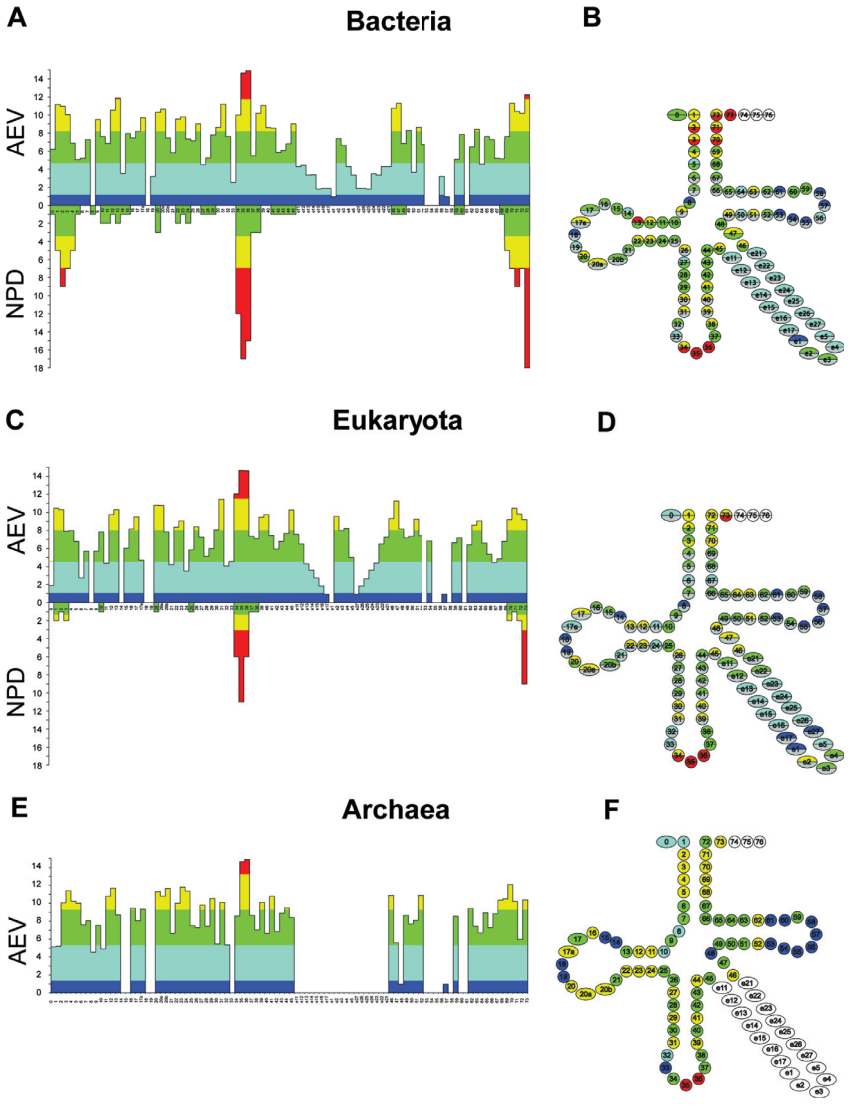
Eredményeimet az élővilág három nagy doménje szerint csoportosítva és ábrázolva (5.8 ábra) mutatom be. Az AEV értékekről statisztikai elemzést készítettem, és azt vizsgáltam, hogy a középértéktől az egyes pozíciók AEV értékei milyen mértékben térnek el. Megvizsgáltam, hogy mely pozíciókban vesz föl az AEV átlagos illetve annál szignifikánsan alacsonyabb és szignifikánsan magasabb értékeket. Az 5.8 ábrán az NPD értékeket is feltüntettem.
Az alábbiakban a bakteriális és eukarióta eredményeket mutatom be, majd – mivel ezeknél a szekvenciáknál a kiindulási adatbázist a *„Módszerfejlesztés"* fejezetben írtak szerint megszűrtem – a bemenő adatok szűrésének hatását mutatom be, végül az ősbakteriális eredményeket ismertetem, ahol nem szűrtem meg a kiindulási adathalmazt.

5.3.2.1. Bakteriális adatok

Az 5.8 A és B ábrán sötétkék színű elemek azok a pozíciók, ahol minden tRNS azonos, tehát az egyes identitások között nem történik megkülönböztetés, ezek a kiindulási adatok a szűréséből adódóan is megjelennek. Ezek a tRNS-ekre jellemző, közös funkciókhoz kellenek, amiket, mint említettem, így a szűrésnél is alkalmaztam. Azok a pozíciók, amelyekre kismértékben jellemzőek a diszkrimináló elemek, átlagos AEV értéket adnak (zölddel jelölve) elsősorban azokra a régiókra esnek, ahol nincsenek ismert identitáselemek. Az akceptor karon: az 5:68-as és 6:67-es pozíció-párokban egy kivétellel még nem írtak le identitáselemeket A kivétel az A5:T68 a Met-re jellemző identitáselem [74].
Az antikodon-hurokban: a konzervált A37-et sok esetben leírták már, mint identitás elemet (Ile

---

5.8. ábra *(lásd túloldalt).* Az AEV és NPD értékei az élővilág három nagy doménjében
A bakteriális (A és B), az eukarióta (C és D) valamint az ősbakteriális szekvenciákra kapott eredmények
Az ECP diagramok
Az egyes oszlopok egy-egy pozíciót jelölnek. Az oszlopok magassága az AEV („átlagos kizárási érték). A diagram színezése statisztikai elemzés eredménye. Az összes pozícióra együttesen vonatkozó eredmény átlaga a zöld terület közepén van. A színezés változása rendre a szórás (szigma) értékeinek +/-0,5 1,5, 2 értékeinél változik. A statisztikai értékek számításakor csak azokat a pozíciókat vettem figyelembe, ahol van (értékelhető) adat. (Nem értékelhető adat a CCA vég – 74-76 pozíció.) A diagram alsó felében az ismert identitáselemek számát (NPD) mutatom be ugyanazokat a színezési elveket alkalmazva az AEV értékeknél használtam.
Az ECP lóherék
Színezése megegyezik a diagram színeivel. Az egyes pozíciókat megjelenítő karikák felső osztásában az AEV értéknek, alsóban az ismert identitáselemek számának (NPD) megfelelő szín van. A figyelembe nem vett illetve 0 értékű pozícióknál a karika szürke.


A
Bacteria
B
AEV
NPD
C
Eukaryota
D
AEV
NPD
E
Archaea
F
AEV

[75], Met [74], Glu [76], Gln [77, 78]), de a tipikus-tRNS ábrán is látszik, hogy sok esetben (a szekvenciák több mint felében) itt A van.
Egyéb régiók: jellemzően nem diszkriminatív pozíciókban nem találunk identitás elemet. Kivételeket jelent: U8:A14 a Leu esetében [79], illetve a Phe-nél a G27:C43, G28:C42 és a T59 [80].

Az AEV alapján leginkább diszkriminatív funkciókat ellátó két elem az antikodon második két tagja (35 és 36 pozíció). Ezek a leggyakoribb identitáselemek is egyben. Megfigyelhető az is, hogy a 34-ik pozíció megkülönböztető szerepe jóval (több mint egy szigmával) kisebb arányú.
A jelentős diszkriminatív szerepű pozíciók nagy része hordoz ismert identitáselemet. Megfigyelhető az is, hogy az egyes Watson-Crick párok elemeinek eredményei korrelálnak (5.8 B ábra).
Az akceptor karon az 1-72 (Trp, Gly, Thr, Gln) a 2-71 (Met, Trp, Asp, Gly, Ser, Cys, Ala, Gln) és a 3-70 (Val, Met, Trp, Gly, Ser, Cys, Ala, Gln) ismert identitáselemek, szinte majdnem mindegyik specifitásban. A diszkriminátor bázis a 73-as pozícióban kiemelt jelentőségű, az antikodon pozíciói után ez rendelkezik a legmagasabb átlagértékkel. A 12-es, jelentősen diszkriminatív pozíció ismert Ile identitáselem: a T12:A23 párban [75]. (A 23-as szintén magas átlagértékkel rendelkezik). Külön érdekesség, hogy a Glu T13:G22:A46 identitáseleméből kettő is (13-as és 46-os) szigmán felüli értékű. A 13:22 pár egyébként még ismert identitáselem a Cys-nél [81, 82]. Ugyanakkor szintén a Glu-nál az említett bázishármason kívül ismert identitáselem a 47 delériója [83–85]. Ezek az identitáselemek a tRNS „magi", avagy „*core*" régiójába esnek [86].
Az átlagosnál magasabb AEV-jű az 5.8 ábrán sárgával jelölt pozíció több ismert identitáselemet tartalmaz: ilyen az antikodon hurokban az 38-as pozíció (Ile, Asp [87, 88], Gln) vagy más helyeken a még nem említett 10-es pozíció (Asp, Gln), a 11:24 (Ser, Glu) és 15:48 pár (Cys, Pro), a 20-as pozíció (Phe, Arg, Ala) és a 29:41 pár (Ile). A 60-as pozícióban, ami egyébként a meglehetősen konzervált T-hurokban található ismert Phe identitáselem van, ahogyan a 45-ösben is [24, 89].
A variábilis hurok leggyakoribb eleme (e2) is diszkriminatív (Ser identitáselem) [90].

Az antikodon huroknál a 31-39 pozíciópárban eddig még nem írtak le identitáselemet. Ugyanakkor az esetek többségében itt egy Watson-Crick pár van. Ennek milyensége azonban identitásfüggő lehet. Az elemzés alapján lehetséges, hogy a 12-23 és a 23-12 pár illetve a 46-os

pozíció az eddig feltárt eseteknél többször vesz részt identitás kialakításában.
Funkcionális jelentőséggel bírhatnak a fakultatív elemek (17, 17a, 20a, 20b) a D-loopban és a 47-es pozícióban. (Lásd még a *„Potenciális identitáselemek"* című fejezetben.)

5.3.3. Eukarióta (élesztő) adatok

Ahogyan a bakteriális adathalmaz esetében a coli-, úgy az eukariótán az élesztő-rokon szekvenciákat vizsgáltam. Eredményeimet a bakteriálissal megegyező módon mutatom be az 5.8 C és D ábrán.

Az átlagosnál alacsonyabb AEV értékek (5.8 C és D ábrán világoskékkel és sötétkékkel) kivétel nélkül a konzervált, identitáselemeket nem tartalmazó pozíciókra esnek. Átlagos AEV értéket mutató (zöld) pozíciókra csak a 3-as esetén esett két identitáselem, a Gly [91] és az Ala [92]. Az antikodon-hurokban található 37-es pozíción csak a Leu [93] a 38-on és a 10-25 bázispáron csak az Asp [94, 95] tartalmaz ismert identitáselemet. A 3-as pozíció esetén párja, a 70-es az átlagosnál magasabb (sárga színnel jelölve) AEV értéket ad.

A legmagasabb (piros) AEV értékeket a legtöbb ismert identitás-elemet hordozó pozíciók, az antikodon bázisai, illetve a diszkriminátor bázis adják. Az akceptor-kar három bázispárja, ahol sok identitáselemet találunk szintén kimagasló AEV értékkel rendelkezik. Kiemelendő ezeken kívül magas AEV értékű (sárga) pozíció, ahol van ismert identitáselem a 20-as, amelyet a Phe esetében írtak le [96].

Magas AEV értékkel találkozhatunk olyan pozíciókban is, ahol még nem írtak le eukarióta identitáselemet. Ezek közül vannak olyanok, amelyet coliban már leírtak, mint identitáselem (lásd még ott: bakteriális adatok eredményei), ilyenek a 12-23, 13-22 bázispárok, amelyek részt vehetnek a *„core régió"* kialakításában a szintén magas AEV értékű 45, 46, 47-es pozíciókkal együtt.

Kiemelendő a 31-39-es és a 30-40-es pár, amelyek nem szerepeltek az adathalmaz szűrési kritériumaként (hiszen ezeken a pozíciókon még nem írtak le élesztő-identitáselemet), azonban humán $tRNS^{Phe}$-nél ezek már ismert indentitás elemek [97].

5.3.4. Az adatszűrés lehetséges hatása az eredményekre

Amint azt a módszerek ismertetésénél leírtam, a felhasznált adatbázisokat háromszorosan szűrtem. A szűrések közül az első és az utolsó magától értetődő lehet. Az elsőben minden tRNS-re közösen jellemző tulajdonságokra szűrtem annak érdekében, hogy valóban csak funkcionális tRNS-eknek megfelelő tDNS szekvenciákkal dolgozzak. A harmadik szűrés egyszerűen csak eltávolította a redundáns szekvenciákat. A második szűrés a logikailag legérdekesebb. Ebben két egymással ellentétes feltétel között igyekeztem kompromisszumos megoldást találni. A tRNS identitás, mint említettem, fajra vonatkozó fogalom, hiszen csak 1-1 fajban kell, hogy együttesen funkcionáljon 20 eltérő identitású tRNS készlet. Így az identitást meghatározó szabályoknak legalább egy része (a legújabban létrejöttek) lehet fajspecifikus. Ez amellett szólna, hogy a szabályok keresésénél kizárólag egy-egy faj tRNS-einek szekvenciáját hasonlítsam össze. Sajnos ez a készlet túl kisszámú bemenő szekvenciát jelent ahhoz, hogy szekvenciában rejlő törvényszerűségek megállapítását lehetővé tegye. A bemenő szekvenciák számát csak úgy lehet növelni, ha növeljük a fajok számát. Ezzel ugyanakkor kérdésessé válik, hogy az így vizsgálatba vont tRNS-ek vajon mind valóban funkcióképesek lennének-e egyetlen közös élőlényben, tehát tökéletesen együttműködnének-e annak szintetáz enzimeivel. Annak érdekében, hogy a bemenő szekvenciák számát növeljem, miközben valószínűsítsem, hogy az egy identitásba tartozó, de eltérő fajból származó tRNS-ek funkcionálisan egyenértékűek legyenek, bevezettem a második szűrést. Ennek során olyan már ismert determinánsok meglétére szűrtem az adathalmazt, amely 1-1 modell élőlényben (*E. coli* illetve élesztő) már leírtak. Úgy érveltem, hogy ezek közös jelenléte az izospecifikus tRNS-ekben növeli annak az esélyét, hogy olyan identitáselemeket is közösen tartalmazzanak, amelyeket kísérletesen még nem tártak fel, de amelyeket a bioinformatikai vizsgálatom valószínűsíthet. Ugyanakkor fontosnak tartottam annak a vizsgálatát, hogy ez a második szűrés torzítja-e, és ha igen miként, az analízis eredményét. Ezért ennek a szűrésnek az elhagyásával is elvégeztem a vizsgálatokat. Mindez azért is fontos volt, mert csak a szűrés nélküli vizsgálat erősítheti meg azoknak a pozícióknak a vizsgálat során kapott szignifikanciáját, amely pozíciók be voltak vonva a szűrésbe. Ezeknél ugyanis kevésbé meglepő, hogy az analízis során is kitűnnek, mint nagy diszkriminációs képességű pozíciók.

A szűrés nélküli adatokon elvégzett statisztikai elemzések eredményeit az 5.3 táblázatban foglaltam össze. A táblázatból kiolvasható, hogy a bakteriális adatoknál, ahol 40 pozíció hordoz ismert identitáselemet, a második szűrési lépés a kiindulási adatok 39%-át eltávolította: azokat, amelyek az adott pozíciókban a megfelelő identitások esetén nem az *E. coli* megfelelő pozíciójában található identitáselemét hordozták. Ennek eredményeképpen azokban a pozíciókban,

5.3. táblázat. A különböző adathalmazok mérete, illetve az elvégzett statisztikai analízisek eredményei

	Bacteria		**Eukaryota**		**Archaea**
	második szűrési lépés nélkül	második szűrési lépéssel	második szűrési lépés nélkül	második szűrési lépéssel	nincs második szűrési lépés
Szekvenciák száma					
Kiindulási adatok	6243		2222		1552
Első szűrési lépés	6144	1930	1384		
Második szűrési lépés	-	3901	-	1672	-
Nem redundáns adatok	3946	2406	1495	1264	1041
AEV					
Átlag	6.45	5.59	5.79	6.31	7.34
Szórás	3.54	3.51	3.43	3.49	3.97
Pearson (R)	0.53	0.55	-	-	-
Spearman (ρ)	0.39	0.54	-	-	-
Bootstrap					
Középérték	224	258	-	-	-
Szórás	16.9	17.1	-	-	-
CAEV küszöb	358.95	344.55	-	-	-
Szignifikancia (P)	1.33e-15	3.54e-7	-	-	-

ahol az NPD pozitív volt, az AEV értéke 22%-kal nőtt, míg azokban a pozíciókban, ahol az NPD 0 volt, ez a növekedés csak 8% volt. A pozitív NPD-jű 40 pozícióban az AEV növekedése 10 esetben haladta meg a szórásának értékét.

Az eukarióta adatokban csak 15 pozitív NPD-jű pozíciói található. A szűrés itt jóval kisebb hatással járt, a szekvenciáknak mindössze a 15%-át távolította el. Ebben az adathalmazban az élesztő identitáselemei alapján szűrtem, amely azt eredményezte, hogy a pozitív NPD-jű pozíciókban az AEV értéke 5%-kal, a 0 NPD-vel rendelkező pozíciókban pedig 9%-kal csökkent. Ez szignifikánsan (szórásnál nagyobb mértékben) csak a 23-as pozícióban csökkent, ahol még nem írtak le identitáselemet.

A bakteriális adathalmazra a statisztikai elemzéseket is elvégeztem úgy is, hogy a kiindulási adatokat nem szűrtem az *E. coli* ismert identitáselemei alapján, és azt kaptam, hogy a szűrés nélküli Pearson korreláció (R) koefficiense 0,55-ről 0,53-ra a Spearman korreláció mértéke (ρ) pedig 0,54-ről 0,39-re csökkent. A *„bootstrap”* analízist is lefuttattam, ennek eredményeit is az 5.3 táblázat tartalmazza.

A statisztikai elemzések mellett a 5.8 ábrán bemutatotthoz hasonlóan a bakteriális és az

eukarióta adathalmaz második szűrési lépés nélküli eredményeit is ábrázoltam a 5.9 ábrán. Az AEV és NPD értékek megállapított korrelációját támasztja alá az, hogy mindkét adathalmaz esetén a legmagasabb AEV értékeket az antikodon bázisai valamint a diszkriminátor bázis kapta. Kiemelkedően magas AEV értékeket kaptam az acceptor-kar terminális bázispárjai esetében is (1:72, 2:71) amelyek szintén sok identitáselemet tartalmaznak szerte az élővilágban.

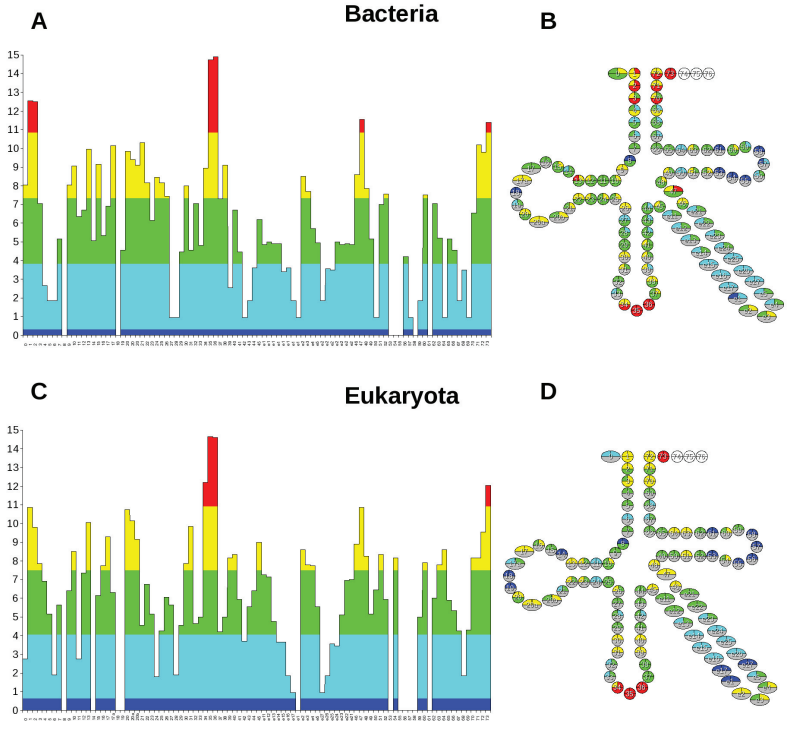
A második szűrési lépés nélküli eredményeimből azonban messzebbmenő funkcionális következtetéseket nem kívánok levonni, hiszen az adatok között szereplő rendkívül sok, fajonként eltérő identitás-készlet alkalmazása miatt általános érvényű megállapításokat – a már ismerteken túl – nem lehet tenni.

5.3.4.1. Az adatszűrés filogenetikai következményei

A második szűrési lépés a bakteriális adathalmaz esetén jól tükrözi az evolúciós viszonyokat. A 1. számú mellékletből kiolvasható, hogy a második szűrési lépés után a legtöbb szekvenciát szolgáltató fajok az *E. coli* legközelebbi rokonai: a Gammaproteobacteria családból (*Escherichia*, *Haemophilus*, *Salmonella*, *Yersinia*, *Buchnera*, *Shigella* nemzettségek) kerülnek ki, itt az első szűrési lépésen már átment szekvenciák legalább 75-85%-a megmarad. Szintén magas (70% körüli) számú szekvencia marad meg a *Proteobacterium*-ok (*Desulfovibrio*, *Brucella*, *Campilobacter*) köréből. Ahogyan a filogenetikai fán egyre távolabb jutunk a második szűrés után megmaradó szekvenciák aránya is csökken, például *Firmicutes* törzs esetén (*Streptococcus*, *Bacillus*, *Lactobacillus*, *Lactococcus*, *Staphylococcus*) 50-70% körül, a *Tenericutes* (*Mycoplasma*, *Ureaplasma*) és *Actinobacteria* törzseknél (*Mycobacterium*, *Streptomyces*) 30-50%.
Hasonló összefüggés az evolúciós kapcsolat és a második szűrési lépés után megmaradó szekvenciák száma között az eukarióta adatok esetén nem figyelhető meg (ennek oka az lehet, hogy az élesztőben kevesebb az NPD, és azok valószínűleg általánosabbak az eukarióták körében).
Fontos megjegyezni, hogy a második szűrési lépés után előállított adathalmaz az egyes fajokból a különböző szűrési szempontok alapján belekerült, egymástól független, különböző identitásokból származó szekvenciák összessége, továbbá azt, hogy amikor coli-szerű és élesztő-szerű adathalmazokra hivatkozom, az a fentiekben ismertetett evolúciós kapcsolatok mellett elsősorban egyfajta (az NPD-k által meghatározott) szekvencia-hasonlóságot jelent.

5.3.5. Ősbakteriális adatok

Az ősbakteriális adatokon identitásonként nem végeztem szűrést. Ennek oka az volt, hogy az irodalomban kevés ősbaktériumra vonatkozó, identitáselemeket feltáró kísérletet jegyeztek fel (összefoglalást lásd 5.4 táblázat). Ezen kívül átfogó, minden aminosav-specifitásra kiterjedő,



5.9. ábra. A bakteriális (A és B) és az eukarióta (C és D) adatok eredményei a második szűrési lépés kihagyásával
A jelölések és a színezés megegyezik a 5.8 ábráéval.
A lóherén az egyes pozíciók felső (AEV-t bemutató) részét kettéosztottam: jobb oldalán a második szűrési lépés kihagyásával, bal oldalán a második szűrési lépéssel született eredményeket mutatom be.

5.4. táblázat. Kísérletesen megállapított ősbakteriális identitáselemek

Aminosav-specifitás	Identitáselemek	Fajok	Irodalom
Ala	G3:U70	*Archaeoglobus fulgidus* *Pyrococcus horikoshii*	[99, 100]
Asp	C36	*Pyrococcus kodakaraensis*	[101–103]
Gly	C35, C36, C2:G71, G3:C70	*Aeropyrum perenix K1*	[104]
His	C73	*Aeropyrum perenix K1*	[105]
Phe	G34, A35 A36, A73, G20	*Aeropyrum perenix K1*	[106]
Pro	G35, G36, A73, G1:C72	*Aeropyrum perenix K1*	[107]
Ser	G30:C40, G73, variábilis hurok G1:C72, C3:G70 variábilis hurok	*Methanosarcina barkeri* *Methanococcus maripaludis*	[108, 109]
Thr	U73, C2:G71	*Haloferax volcanii* *Aeropyrum perenix K1*	[110–112]
Trp	C34, C35, A36, A73, G1:C72, G2:C71	*Aeropyrum perenix K1*	[113]
Tyr	C1:G72, A73	*Aeropyrum perenix K1*	[114, 115]

identitáselemeket feltáró munkát kizárólag in silico, szekvencia-illesztés alapján végeztek [98].

Az ősbaktériumokra kapott legalacsonyabb AEV értékek a bakteriális adathalmazhoz hasonlóan a konzervált pozíciókban vannak. Szembetűnő különbség a bakteriális AEV értékekhez képest az, hogy az 1-72 bázispár az akceptor karon átlagos (5.8 D és E ábrán zölddel jelölve) értékeket vesz föl. Ez annak az eredménye lehet, hogy az ősbaktériumok acceptor-karján az esetek több mint 90%-ban G1:C72 párt találunk, a szekvencia-elemzés alapján [98] kivételt a Tyr C1:G72 bázispárja jelent, amely identitáselem mivoltát kísérlettel is igazolták [114, 115]. A szekvencia-elemzések az iMet (iniciátor) és Gln szekvenciákat is ettől eltérőnek mutatták.

Az antikodon 34. pozíciója a bakteriális adatokhoz hasonlóan alacsonyabb AEV értéket ad a 35-36-osnál, azonban itt, az ősbaktériumoknál az átlagos (zöld) tartományba került.

Magas (az átlagosnál nagyobb, sárga ill. piros színnel jelölve a 5.8 D és E ábrákon) AEV értékeket kaptak az ismert identitáselemeket tartalmazó pozíciók, az antikodon tagjai és a diszkriminátor bázis. Ezeket kísérletesen is több aminosav identitás esetén is sikerült már megállapítani (lásd 5.4 táblázat).

Ugyanakkor az Thr RNS példáján keresztül például láthatjuk, hogy az élővilág egyes csoportjai között, illetve azokon belül hogyan különbözhetnek az AC-karon lévő egyes identitáselemek. Az *E. coli* esetén [116] a diszkriminátor bázis nem játszik szerepet, míg az AC-kar három bázispárja (leginkább a második, 2-71) szerepet játszik a felismerésben. Az élesztő esetén [117] a diszkriminátor valamint az első és harmadik bázispár bír identitás szereppel, és a *Thermus thermophilus* baktériumnál is az élesztőhöz hasonló eredményre jutottak [118]. Két ősbakterium

faj Thr tRNS-t vizsgáló kísérlet során [110] rámutattak arra, hogy hasonló eltérés tapasztalható a diszkriminátor bázis és az AC-kar különböző bázispárjai között: a Haloferax volcanii az élesztőre és a *T. thermophilus*-ra, az és *Aeropyrum perenix* az *E. coli*-ra hasonlított ebben a tekintetben. A kapott AEV értékek arra engednek következtetni, hogy az ősbaktériumoknál az *A. perenix*-hez hasonló identitáselem-mintázattal találkozhatunk, illetve a diszkriminátor bázis jelentősége akár több más identitásban is kisebb mértékű lehet. Emellett a 3-70 bázispár jelentőségét is több identitás esetén (Ala, Gly, Ser: lásd 5.4 táblázat) leírták, és ez a bázispár szintén kiemelkedően magas AEV értéket kapott.
Az átlagosnál magasabb AEV értékkel (sárga) rendelkezik még a 20-as pozíció, amely az ősbaktériumoknál is leírt Phe identitáselem [106].

Az antikodon hurok karján lévő 29-41 ill. 31-39 bázispárok is átlagosnál magasabb AEV értékűek, azonban itt nem írtak le még identitáselemet, csak az átlagos, illetve az alatti tartományba eső 30-40 bázispárban találtak Ser identitáselemet [108] a *Methanosarcina barkeri* egyik szintetáza esetében.
A bakteriális AEV értékekhez hasonlóan az átlagosnál magasabb, még leírt identitáselemet nem tartalmazó pozíciók a *„core régió"*-ban és pl. különböző fakultatív bázisokat tartalmazó pozíciókban vannak (lásd még 5.8 E ábra).

5.3.6. Potenciális identitáselemek

A lehetséges, eddig föl nem tárt identitáselemek kiválasztásakor azok a pozíciók jöhetnek elsősorban szóba, amelyeknek magas AEV értékük van, ugyanakkor nem tartalmaznak ismert identitáselemet, illetve a pozícióban az adatok szűrésekor nem vettem figyelembe már ismert identitáselemet, mint szűrési szempontot. Ezeket a pozíciókat megvizsgálva olyan irodalmi adatok után kutattam, amelyek – ha coli illetve élesztő esetére nem is közöltek identitáselemeket, de esetleg más fajokra vonatkozóan közöltek adatokat. Ilyen esetnek adódott a már említett 31-39 és 30-40 bázispár, amely a humán Phe-nál identitáselem (Nazarenko 1992). Ezért e két bázispárt megvizsgáltam a kiindulási szekvenciák összes aminosav-identitású tDNS-e között. A legkézenfekvőbb az vol, hogy olyan, egyedi aminosav-specifitást keressek, amelynél a két említett bázispár közül legalább az egyik minden más aminosavidentitású tRNS-készlettől eltér. Eukarióta (élesztő) szekvenciákban ezek az azonosított esetek kivétel nélkül nem Watson-Crick („*wobble*") bázispárok.

5.3.6.1. *E. coli* Trp T31:A39

A coli szekvenciák közül négyből egy Ser, illetve kettőből egy Gln szekvencia mutat hasonló mintázatot. A Trp identitást coliban eddig csak az antikodon-hurokban illetve diszkriminátor pozícióban és az AC-karon vizsgáltak részletesen [119–121]. A T31:A39 bázispárral rendelkező Trp-os szekvenciák a coli-rokon (szűrt adatbázisunkban szereplő) fajokban – egy-két kivételtől eltekintve – megtalálhatóak.

5.3.6.2. Élesztő Met T31:T39

Ez a nem Watson-Crick bázispár az eukarióta fajok „elongátor" $tRNS^{Met}$ molekulájára jellemző. Az eukarióta és *E. coli* „iniciátor" $tRNS^{Met}$ valamint az *E. coli* „elongátor" $tRNS^{Met}$ molekulája ugyanitt normális Watson-Crick, mégpedig G31:C39 bázispárt tartalmaz. Az *E. coli* „iniciátor" $tRNS^{Met}$-ben a normális G31:C39 bázispár ahhoz szükséges, hogy a fehérjeszintézis kezdőlépésekor a riboszóma P-helyéhez tudjon kötődni a tRNS, az *E. coli* „elongátorban" pedig ahhoz, hogy a megfelelő aaRS helyesen töltse föl a tRNS-t. Az eukarióta „iniciátor" tRNS az *E. coli* Met szintetáz enzimnek (amellyel a természetben soha nem találkozik) jó szubsztrátja, míg az eukarióta „elongátor" $tRNS^{Met}$ a coli enzim számára nem jó szubsztrát (Pawel-Rammingen 1992; Drabkin 1993). Amikor az eukarióta elongátor $tRNS^{Met}$-ben lévő eredeti T31:T39 bázispárt kicserélték G31:C39 párra, az jó szubsztrátnak bizonyul az *E. coli* Met szintetáz enzime számára, és szimmetrikusan, mikor az *E. coli* „elongátor" $tRNS^{Met}$-ben az eredeti G31:C39 párt T31:T39 párra cserélték, az jó szubsztrát lett az eukarióta enzim számára. (Meinnel 1992) Ezek a *„kingdom"*-ra jellemző bázispárok tehát mindkét *„kingdom"*-ban identitáselemekként viselkednek. Az is bizonyítást nyert, hogy nem maguk a bázispárban lévő bázisok számítanak, hanem az, hogy ezek Watson-Crick párt alkotnak-e, vagy sem. Ha Watson-Crick párt alkotnak, akkor rossz szubsztrátjai lesznek az eukarióta és jó szubsztrátjai a bakteriális enzimnek. Ha nem alkotnak Watson-Crick párt, akkor fordítva, jó szubsztrátjai lesznek az eukarióta, és rossz szusztrátjai a bakteriális enzimnek. Ebben a pozícióban az, hogy jelen van-e vagy éppen nincs-e jelen Watson-Crick bázispár befolyásolja az antikodon hurok szerkezetét és/vagy deformálhatóságát, ami fontos szerepet játszhat a megfelelő szintetázzal való kölcsönhatásban. Azt is kimutatták, hogy amennyiben az élesztő iniciátor $tRNS^{Met}$-ben kicserélik a Watson-Crick G31:C39 bázispárt T31:T39 párra, úgy a molekula képessé válik arra, hogy elongátor tRNS-ként működjön. Mindezek ellenére ezt a bázispárt az irodalomban mégsem definiálják determinánsként, ugyanis ezt az elnevezést csak akkor érdemli ki egy pozíció, ha azonos fajban cserélik ki két vagy több elongátor tRNS között a vizsgált részeket. A jelen esetben eltérő *„kingdom"*-ok elongátor tRNS-ei között zajlottak ezek a cserék, illetve azonos faj esetén elongátor és iniciátor tRNS-ek között.

5.3.6.3. Élesztő Ile T30:G40

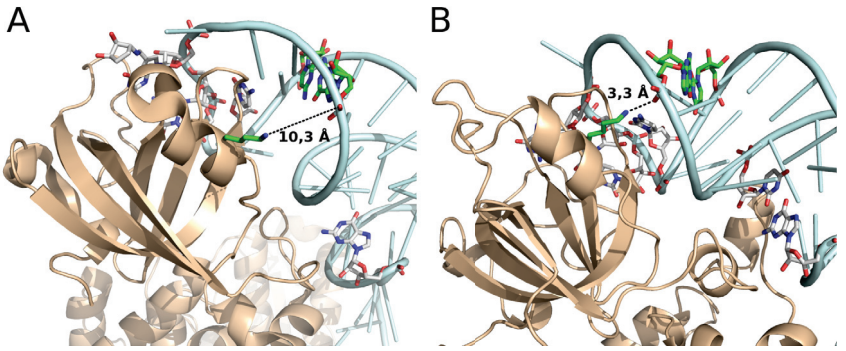
Az élesztő Ile tRNS identitáselemeit az antikodon-bázishármasában vizsgálták, ahol rámutattak a módosított bázisok szerepére (Senger 1997). Az itt jelzett rendhagyó bázispár szerepét még nem vizsgálták. Adatbázisunkban az eukarióta fajok többségénél (az arabidobsistól a muslincán át az emberig) megtalálható, több Ile izoakceptor esetén is.

5.3.6.4. Élesztő Asp G30:T40

Szintén rendhagyó bázispárosodást mutat az élesztő Asp tRNS-e ugyanabban a pozícióban, mint az Ile. Az Ala és Phe esetén már régen leírták [122, 123] ennek szerepét az identitásban, illetve később ennek szerkezeti okait is feltárták (Chang 1999). Az Asp-ban már van rendhagyó bázispár, mint identitáselem: a G10:U25 bázispár [94, 95]. A G30:T40 rendhagyó bázispár ugyanakkor nem elterjedt az eukarióták között, adatbázisunkban mindössze az élesztőben és a C. elegansban található meg.
A magas AEV érték alapján potenciális identitáselemet feltételeztem itt, amelynek szerkezeti okait kutatva megvizsgáltam az ismert tRNSAsp – AspRS térszerkezet [124]. A szerkezetet leíró közleményben megállapították, hogy a háromdimenziós szerkezet sokkal fontosabb szerepet játszik a tRNS-szintetáz kapcsolatban, mint egy-egy identitáselem.

Ezután a PDB adatbázisban hozzáférhető élesztő aaRS-tRNS complex szerkezeteket tanulmányoztam, különös tekintettel a 30:40 bázispár szerkezetbeli környezetére. Feltételezhető ionos kölcsönhatásra következtethetünk a tRNSAsp és a szintetáz esetében a G30 cukor-foszfát gerince és a Lys88 oldallánca között, távolságuk a kristályban található két szerkezetben 2,9 Å és 3,3 Å. A másik két szerkezet (Tyr, Arg) közül a Tyr (PDB: 2DLC) hiányos (a kristály nem szórt megfelelően). Az Arg [1] esetében a C40 cukor-foszfát gerinc a Ser440-nel alkothat hidrogénhidat (távolságuk 2,6 Å). A szintetázok szekvenciáinak MUSCLE-illesztésével az Asp-aaRS Lys88 csoportjának homológ megfelelője, az Arg-aaRS Lys78 csoportja nem vesz részt a tRNS-szintetáz interakcióban. Ezen kívül megmértem a 30:40 (a tRNSAsp –nál ez G:T, a többi esetben G:C) cukor-foszfát gerinceinek távolságait. A $tRNS^{Asp}$ –nál ez rendre 18,9 Å és 19,7 Å, a $tRNS^{Arg}$ –nál 18,8 Å, a $tRNS^{Tyr}$ –nál pedig 17,7 Å.

Hipotézisemre, miszerint az élesztőben (és a tRNS szekvenciák egyezése miatt feltehetően *C. elegans*-ban is) a G30:T40 bázispár identitáselem, szerkezeti bizonyítékokat is próbáltam találni, mivel az Asp tRNS és az Asp szintetáz komplexe mind *E. coli*-ban [125] mind pedig élesztőben [124] ismert. Az élesztőben a G:T bázispárból a G30 cukor-foszfát gerincével a

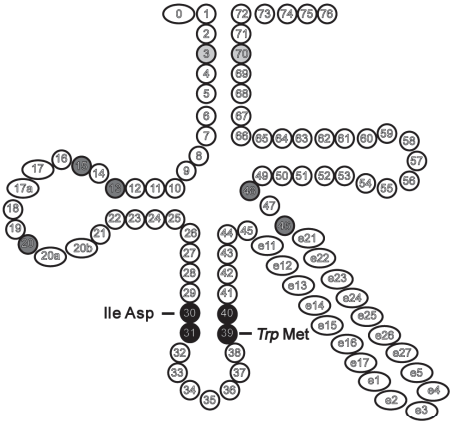


5.10. ábra. tRNSAsp – AspRS komplex szerkezetek
A) *E. coli* tRNSAsp és AspRS szerkezete [125]; PDB: 1IL2
B) élesztő tRNSAsp és AspRS szerkezete [124]; PDB: 1ASY
színezések: aaRS: halvány barna; tRNS: szürkéskék; zölddel a Lys58 és a Lys155- öt, illetve a G30:C40 és G30:U40 bázispárokat emeltem ki, a többi pálcikával ábrázolt bázis az ismert identitáselemek.

szintetáz Lys155 nagy valószínűséggel ionos kölcsönhatást tud képezni (távolságuk 3,3 Å , lásd 5.10 B ábra), amely stabilizálhatja a komplex szerkezetét. Ezután az élesztő és *E. coli* szintetázok szekvenciáit Needleman-Wunsch algoritmussal illesztve megállapítottam, hogy az *E. coli*-ban a Lys155-tel a Lys58 homológ (a 5.10 B ábrán szintén kiemelve). A kérdéses foszfát-gerincől ez már jóval távolabb 10,3Å -re helyezkedik el. Feltételezésem szerint a szintetázok kismértékben eltérő szerkezete mellett az élesztő tRNS-ben lévő G:T pár a tRNS szerkezetét jobban „lazítja", mint egy szabályos, Watson-Crick G: C bázispár, amely kedvez a szintetázzal való „sóhíd", a stabilabb tRNS-aaRS szerkezet kialakulásának.

5.3.6.5. *„Core-régió"*

A core-régióra eső magas AEV értékekből arra lehet következtetni, hogy a már leírt *E. coli* Pro-nál [126] és Cys-nél (G15:G48, (Hou 1999) valamint a Glu identitásban [85] betöltött szerepén kívül (megkülönböztetés Asp-tól) az eukarióta illetve az ősbakteriális fajok esetében (Ryckelynck 2003) is szerepet játszhat.
A potenciális identitáselemeket összefoglalva ábrázoltam a 5.11 ábrán.



5.11. ábra. Lehetséges, eddig nem ismert identitáselemek

Azokat a pozíciókat jelöli az ábra, amelyek alacsony NPD értékkel, de magas AEV-vel rendelkeznek (szürkével és feketével jelölve). A feketével jelölt potenciális identitáselemek közül azt az aminosavspecifitást, amely az *E. coli*-ban feltételezhető dőlt betűvel, azok, amelyek az élesztőben jósolhatóak, vastag betűkkel jelöltek. Sötétszürkével kiemeltek a „*core régió*" mindhárom élőlénycsoportban magas AEV-jű pozíciói. Világos szürkével ábrázolt az ősbakteriális szekvenciák legmagasabb AEV-jű eredményei.

5.4. Konklúzió

Az új, diszkrét matematikai módszer, az ECP kifejlesztése és lehetőségeinek kibővítése arra irányult, hogy a tRNS-ek identitáselemeit minél alaposabban feltárhassuk. Törekvésem az volt, hogy a meglehetősen hosszú, de még számtalan kérdést magában rejtő [127] témához egy új bioinformatikai eszközzel járuljak hozzá. Olyannal, amely az alkalmazott logika tekintetében lényegesen eltér az eddigiektől, ezért esetleg lehetővé teszi olyan új identitáselemek előrejelzését, amelyek eddig nem kerültek a figyelem középpontjába. Noha eddig is léteztek már in silico identitás előrejelző eszközök, ezek azokra az elemekre fókuszáltak, amelyek konzervált szekvencia-motívumokként jelentek meg a tRNS szekvenciákban. Merőben új volt Jakó Éena megközelítése, amely elsősorban a hiányzó elemekre koncentrált. Ezt a megközelítést alkalmaztam és fejlesztettem tovább munkám során.

Az ECP működése kapcsán bemutattam, hogy a módszer nem egyedi szekvenciákat, hanem azok csoportjait (legyen szó osztályokról vagy identitásokról), a csoportok egymástól való távolságát adja meg. Ez a távolság minimális akkor, ha egy pozícióban minden szekvencia esetén ugyanazt a nukleotidot tartalmazza mindkét csoport mindegyik szekvenciája, de ugyanúgy minimális akkor is, ha az egyik illetve másik csoport szekvenciái között előfordul mind a négyféle nukleotid. Ezekben az esetben nem kapunk DE-et. Ha az identitáselemek szempontjából kívánjuk ezt a jelenséget magyarázni, arra következtethetünk, hogy ezekben a pozíciókban, ahol minimális az ECP-távolság, a csoportok egymáshoz a legközelebb állnak, ott az evolúció – a csoportok megkülönböztetése szempontjából – megengedő volt. Ez két esetben lehetséges. Egyrészt azért, mert a csoportok (identitások) ebben a pozícióban nem kell, hogy megkülönböztessék egymást (a teljesen véletlenszerűen előforduló nukleotidok esete), így szabadon mutálódhatnak, nincsen rajtuk evolúciós nyomás. A másik eset az, hogy ez a pozíció valamilyen más funkcióra fenntartott elemet tartalmaz (a mindkét csoportban minden szekvenciában ugyanazt a nukleotidot tartalmazó pozíció esete), amelynek szerepe lehet olyan közös tRNS funkciók ellátásában, mint például a riboszómához való kötődés.

Az első alkalmazás előrelépést jelentett egy régi dogma [22] megcáfolásában: hatékony eszközzé válhatott két szekvencia-csoport eddig nem tapasztalt mértékű szétválasztásában és egy olyan, a két szétválasztott osztályra jellemző, specifikus adathalmaz létrehozásában, amely az eredeti szekvenciákból hiányzó elemeken alapul. Az identitásokat egymással szemben, egy-egy ilyen csoportnak (osztálynak) tekintve, egymással kombinálva (a 20 aminosavat a 19 másikkal párba állítva) pedig alkalmassá lehet tenni ezt a módszert az identitáselemek „forró pontjainak”

feltérképezéséhez.

A szekvencia-analízisen alapuló munkák egyik fő igénye az, hogy a bemenő adatok megfelelően sok információval szolgálhassanak megbízható, pontos következtetések levonásához. Amint azt korábban már említettem, még bio-statisztikai szemmel nézve is kevés adatot produkál egy-egy faj tRNS készlete, még az eukarióta fajok közt is. Éppen ezért kellett munkám során olyan evolúciósan rokon szekvenciákhoz nyúlnom, amelyek a bemenő adatok számát megnövelik úgy, hogy emellett vélhetően funkcionálisan is releváns információkkal tudnak szolgálni. Mindez azon a feltételezésen alapult, hogy az identitásonként elvégzett, bizonyos ismert identitáselemek meglétén alapuló szűrés olyan tRNS készleteket eredményezhet, amelyek elemei működhetnek abban a fajban, amelyből a szűréshez használt szabályok származtak.

A nagymennyiségű adat ugyanakkor mégis vezethet ahhoz, hogy fals pozitív elemekkel szennyezzük az adatkészletet. Olyan tRNS-ekkel, amelyek mégsem működnének a szűréshez használt modell fajban. Hiszen, mint említettem, tRNS identitáselemek magától értetődő módon egy-egy faj esetében értelmezhetőek, a természetben egy-egy faj tRNS-e nem találkozik a másik faj szintetázával. Az evolúció során az identitáselemek akár véletlen sodródással is eltérővé válhattak az egymástól elváló fajokban. Fontos gyakorlati kérdés, hogy ezek az eltérések akár új antibiotikumok kifejlesztését is lehetővé tehetik az által, hogy az eberi és a kórokozó rendszer közötti eltéréseket feltárják.

A fenti eszmefuttatás mentén a következő logikát alkalmaztam: a bakteriális adatokat az *E. coli*, az eukariótát az élesztő ismert identitáselemei alapján szűrtem meg. Gondolatmenetem az volt, hogy a szűrés után az evolúciósan rokon szekvenciák maradnak meg. Emiatt feltételeztem, hogy nem csak az ismert identitáselemeik lesznek közösek, de azok is, amelyeket nem használtam a szűréshez, illetve amelyeket eddig még nem is tártak fel. Ez természetesen óhatatlanul azzal is járt, hogy azokat az identitáselemeket az analízisem nem tárja majd fel, amelyek szigorúan csak egy-egy fajra jellemzőek.

Statisztikai módszerekkel bebizonyítottam azonban, hogy az adatok ilyen szűrése az AEV alapvető karakterisztikáját nem befolyásolják: az eljárás a szűrés elhagyásával is azt eredményezi, hogy a magas AEV értékek zömmel ismert, tehát valódi identitáselemeket hordozó pozíciókon jelennek meg. Az AEV értékek tehát szűrés nélkül is korrelálnak az ismert identitáselemek előfordulási gyakoriságával. Ezzel a továbbfejlesztéssel az ECP algoritmus immár képessé vált arra, hogy kibővített adatbázisokon akár identitásonként tárja fel a diszkrimináló (más identitást kizáró) pozíciókat.

Mindemellett nem pusztán arra törekedtem, hogy konkrét tRNS identitások egy-egy bázisát/bázispárját azonosítsam potenciális identitáselemként. Ehelyett olyan „forró pontokat” pró-

báltam feltérképezni, amelyek - bár az élővilág nagyobb csoportjaiban hordozhatnak identitás elemeket-, mégsem kerültek eddig a részletesebb vizsgálatok keresztttüzébe.
Eredményeimet azzal a fenntartással kezelem, hogy azok akkor nyernek majd igazán értelmet, ha a kísérleti munkák alátámasztják relevanciájukat.

Irodalomjegyzék

[1] B. Delagoutte, D. Moras, and J. Cavarelli, „trna aminoacylation by arginyl-trna synthetase: induced conformations during substrates binding." *EMBO J*, vol. 19, no. 21, pp. 5599–5610, Nov 2000. [Online]. Available: http://dx.doi.org/10.1093/emboj/19.21.5599

[2] E. Freyhult, V. Moulton, and D. H. Ardell, „Visualizing bacterial trna identity determinants and antideterminants using function logos and inverse function logos." *Nucleic Acids Res*, vol. 34, no. 3, pp. 905–916, 2006. [Online]. Available: http://dx.doi.org/10.1093/nar/gkj478

[3] S. A. Martinis, P. Plateau, J. Cavarelli, and C. Florentz, „Aminoacyl-trna synthetases: a family of expanding functions. mittelwihr, france, october 10-15, 1999." *EMBO J*, vol. 18, no. 17, pp. 4591–4596, Sep 1999. [Online]. Available: http://dx.doi.org/10.1093/emboj/18.17.4591

[4] P. Mucha, „Aminoacyl-trna synthetases and aminoacylation of trna in the nucleus." *Acta Biochim Pol*, vol. 49, no. 1, pp. 1–10, 2002.

[5] C. Carter, Jr, „Cognition, mechanism, and evolutionary relationships in aminoacyl-trna synthetases." *Annu Rev Biochem*, vol. 62, pp. 715–748, 1993. [Online]. Available: http://dx.doi.org/10.1146/annurev.bi.62.070193.003435

[6] J. Cavarelli and D. Moras, „Recognition of trnas by aminoacyl-trna synthetases." *FASEB J*, vol. 7, no. 1, pp. 79–86, Jan 1993.

[7] S. Cusack, „Aminoacyl-trna synthetases." *Curr Opin Struct Biol*, vol. 7, no. 6, pp. 881–889, Dec 1997.

[8] R. Giegé, „The early history of trna recognition by aminoacyl-trna synthetases." *J Biosci*, vol. 31, no. 4, pp. 477–488, Oct 2006.

[9] P. Schimmel, R. Giegé, D. Moras, and S. Yokoyama, „An operational rna code for amino acids and possible relationship to genetic code." *Proc Natl Acad Sci U S A*, vol. 90, no. 19, pp. 8763–8768, Oct 1993.

[10] M. Szymański, M. Deniziak, and J. Barciszewski, „The new aspects of aminoacyl-trna synthetases." *Acta Biochim Pol*, vol. 47, no. 3, pp. 821–834, 2000.

[11] S. Cusack, M. Härtlein, and R. Leberman, „Sequence, structural and evolutionary relationships between class 2 aminoacyl-trna synthetases." *Nucleic Acids Res*, vol. 19, no. 13, pp. 3489–3498, Jul 1991.

[12] G. Eriani, M. Delarue, O. Poch, J. Gangloff, and D. Moras, „Partition of trna synthetases into two classes based on mutually exclusive sets of sequence motifs." *Nature*, vol. 347, no. 6289, pp. 203–206, Sep 1990. [Online]. Available: http://dx.doi.org/10.1038/347203a0

[13] G. M. Nagel and R. F. Doolittle, „Evolution and relatedness in two aminoacyl-trna synthetase families." *Proc Natl Acad Sci U S A*, vol. 88, no. 18, pp. 8121–8125, Sep 1991.

[14] C. R. Woese, G. J. Olsen, M. Ibba, and D. Söll, „Aminoacyl-trna synthetases, the genetic code, and the evolutionary process." *Microbiol Mol Biol Rev*, vol. 64, no. 1, pp. 202–236, Mar 2000.

[15] M. Ibba, S. Morgan, A. W. Curnow, D. R. Pridmore, U. C. Vothknecht, W. Gardner, W. Lin, C. R. Woese, and D. Söll, „A euryarchaeal lysyl-trna synthetase: resemblance to class i synthetases." *Science*, vol. 278, no. 5340, pp. 1119–1122, Nov 1997.

[16] M. Ibba, A. W. Curnow, and D. Söll, „Aminoacyl-trna synthesis: divergent routes to a common goal." *Trends Biochem Sci*, vol. 22, no. 2, pp. 39–42, Feb 1997.

[17] M. Ibba, J. L. Bono, P. A. Rosa, and D. Söll, „Archaeal-type lysyl-trna synthetase in the lyme disease spirochete borrelia burgdorferi." *Proc Natl Acad Sci U S A*, vol. 94, no. 26, pp. 14 383–14 388, Dec 1997.

[18] D. Söll, H. D. Becker, P. Plateau, S. Blanquet, and M. Ibba, „Context-dependent anticodon recognition by class i lysyl-trna synthetases." *Proc Natl Acad Sci U S A*, vol. 97, no. 26, pp. 14 224–14 228, Dec 2000. [Online]. Available: http://dx.doi.org/10.1073/pnas.97.26.14224

[19] M. Ibba, H. C. Losey, Y. Kawarabayasi, H. Kikuchi, S. Bunjun, and D. Söll, „Substrate recognition by class i lysyl-trna synthetases: a molecular basis for gene displacement." *Proc Natl Acad Sci U S A*, vol. 96, no. 2, pp. 418–423, Jan 1999.

[20] T. Terada, O. Nureki, R. Ishitani, A. Ambrogelly, M. Ibba, D. Söll, and S. Yokoyama, „Functional convergence of two lysyl-trna synthetases with unrelated topologies." *Nat Struct Biol*, vol. 9, no. 4, pp. 257–262, Apr 2002. [Online]. Available: http://dx.doi.org/10.1038/nsb777

[21] T. Brennan and M. Sundaralingam, „Structlre of transfer rna molecules containing the long variable loop." *Nucleic Acids Res*, vol. 3, no. 11, pp. 3235–3250, Nov 1976.

[22] H. Nicholas, Jr and W. H. McClain, „Searching trna sequences for relatedness to aminoacyl-trna synthetase families." *J Mol Evol*, vol. 40, no. 5, pp. 482–486, May 1995.

[23] W. H. McClain, „Rules that govern trna identity in protein synthesis." *J Mol Biol*, vol. 234, no. 2, pp. 257–280, Nov 1993. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1993.1582

[24] R. Giegé, M. Sissler, and C. Florentz, „Universal rules and idiosyncratic features in trna identity." *Nucleic Acids Res*, vol. 26, no. 22, pp. 5017–5035, Nov 1998.

[25] J. M. Sherman and D. Söll, „Aminoacyl-trna synthetases optimize both cognate trna recognition and discrimination against noncognate trnas." *Biochemistry*, vol. 35, no. 2, pp. 601–607, Jan 1996. [Online]. Available: http://dx.doi.org/10.1021/bi951602b

[26] M. Hoagland, „Biochemistry or molecular biology? the discovery of 'soluble rna'." *Trends Biochem Sci*, vol. 21, no. 2, pp. 77–80, Feb 1996.

[27] A. Rich, *Horizons In Biochemistry. Kasha M., Pullman B. , editors. New York: Academic Press; 1962. p. 103-126*, B. P. M. Kasha, Ed. New York: Academic Press, 1962.

[28] P. Lengyel, „Problems in protein biosynthesis." *J Gen Physiol*, vol. 49, no. 6, pp. 305–330, Jul 1966.

[29] H. Matthaei, „Proceedings of the mendel centenial symposium." in *The University of Wisconsin Press*, 1965.

[30] M. J. Rogers, T. Adachi, H. Inokuchi, and D. Söll, „Switching trna(gln) identity from glutamine to tryptophan." *Proc Natl Acad Sci U S A*, vol. 89, no. 8, pp. 3463–7, Apr. 1992.

[31] M. Ibba, H. C. Losey, Y. Kawarabayasi, H. Kikuchi, S. Bunjun, and D. Söll, „Substrate recognition by class i lysyl-trna synthetases: a molecular basis for gene displacement." *Proc Natl Acad Sci U S A*, vol. 96, no. 2, pp. 418–23, Jan. 1999.

[32] S. A. Martinis and P. Schimmel, „Microhelix aminoacylation by a class i trna synthetase. non-conserved base pairs required for specificity." *J Biol Chem*, vol. 268, no. 9, pp. 6069–72, Mar. 1993.

[33] K. Nakanishi, S. Fukai, Y. Ikeuchi, A. Soma, Y. Sekine, T. Suzuki, and O. Nureki, „Structural basis for lysidine formation by atp pyrophosphatase accompanied by a lysine-specific loop and a trna-recognition domain." *Proc Natl Acad Sci U S A*, vol. 102, no. 21, pp. 7487–92, May 2005.

[34] J. ichi Fukunaga, S. Ohno, K. Nishikawa, and T. Yokogawa, „A base pair at the bottom of the anticodon stem is reciprocally preferred for discrimination of cognate trnas by escherichia coli lysyl- and glutaminyl-trna synthetases." *Nucleic Acids Res*, vol. 34, no. 10, pp. 3181–8, 2006.

[35] R. L. Sherrer, J. M. L. Ho, and D. Söll, „Divergence of selenocysteine trna recognition by archaeal and eukaryotic o-phosphoseryl-trnasec kinase." *Nucleic Acids Res*, vol. 36, no. 6, pp. 1871–80, Apr. 2008.

[36] N. Nameki, „Identity elements of trna(thr) towards saccharomyces cerevisiae threonyl-trna synthetase." *Nucleic Acids Res*, vol. 23, no. 15, pp. 2831–6, Aug. 1995.

[37] V. Büttcher, B. Senger, S. Schumacher, J. Reinbolt, and F. Fasiolo, „Modulation of the suppression efficiency and amino acid identity of an artificial yeast amber isoleucine transfer rna in escherichia coli by a g-u pair in the anticodon stem." *Biochem Biophys Res Commun*, vol. 200, no. 1, pp. 370–7, Apr. 1994.

[38] B. Senger, S. Auxilien, U. Englisch, F. Cramer, and F. Fasiolo, „The modified wobble base inosine in yeast trnaile is a positive determinant for aminoacylation by isoleucyl-trna synthetase." *Biochemistry*, vol. 36, no. 27, pp. 8269–75, July 1997.

[39] T. Muramatsu, K. Nishikawa, F. Nemoto, Y. Kuchino, S. Nishimura, T. Miyazawa, and S. Yokoyama, „Codon and amino-acid specificities of a transfer rna are both converted by a single post-transcriptional modification." *Nature*, vol. 336, no. 6195, pp. 179–81, Nov. 1988.

[40] K. Tamura, H. Himeno, H. Asahara, T. Hasegawa, and M. Shimizu, „In vitro study of e.coli trna(arg) and trna(lys) identity elements." *Nucleic Acids Res*, vol. 20, no. 9, pp. 2335–9, May 1992.

[41] K. Breitschopf, T. Achsel, K. Busch, and H. J. Gross, „Identity elements of human trna(leu): structural requirements for converting human trna(ser) into a leucine acceptor in vitro." *Nucleic Acids Res*, vol. 23, no. 18, pp. 3633–7, Sept. 1995.

[42] T. Suzuki, T. Ueda, and K. Watanabe, „The 'polysemous' codon–a codon with multiple amino acid assignment caused by dual specificity of trna identity." *EMBO J*, vol. 16, no. 5, pp. 1122–34, Mar. 1997.

[43] J. Pütz, C. Florentz, F. Benseler, and R. Giegé, „A single methyl group prevents the mischarging of a trna." *Nat Struct Biol*, vol. 1, no. 9, pp. 580–2, Sept. 1994.

[44] A. Fender, R. Geslain, G. Eriani, R. Giegé, M. Sissler, and C. Florentz, „A yeast arginine specific trna is a remnant aspartate acceptor." *Nucleic Acids Res*, vol. 32, no. 17, pp. 5076–86, 2004.

[45] A. Soma, R. Kumagai, K. Nishikawa, and H. Himeno, „The anticodon loop is a major identity determinant of saccharomyces cerevisiae trna(leu)." *J Mol Biol*, vol. 263, no. 5, pp. 707–14, Nov. 1996.

[46] D. H. Ardell, „Computational analysis of trna identity." *FEBS Lett*, vol. 584, no. 2, pp. 325–333, Jan 2010. [Online]. Available: http://dx.doi.org/10.1016/j.febslet.2009.11.084

[47] T. M. Lowe and S. R. Eddy, „trnascan-se: a program for improved detection of transfer rna genes in genomic sequence." *Nucleic Acids Res*, vol. 25, no. 5, pp. 955–964, Mar 1997.

[48] S. R. Eddy and R. Durbin, „Rna sequence analysis using covariance models." *Nucleic Acids Res*, vol. 22, no. 11, pp. 2079–2088, Jun 1994.

[49] D. Laslett and B. Canback, „Aragorn, a program to detect trna genes and tmrna genes in nucleotide sequences." *Nucleic Acids Res*, vol. 32, no. 1, pp. 11–16, 2004. [Online]. Available: http://dx.doi.org/10.1093/nar/gkh152

[50] D. H. Gauss, F. Grüter, and M. Sprinzl, „Compilation of trna sequences." *Nucleic Acids Res*, vol. 6, no. 1, pp. r1–r19, Jan 1979.

[51] M. Sprinzl and K. S. Vassilenko, „Compilation of trna sequences and sequences of trna genes." *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D139–D140, Jan 2005. [Online]. Available: http://dx.doi.org/10.1093/nar/gki012

[52] F. Jühling, M. Mörl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Pütz, „trnadb 2009: compilation of trna sequences and trna genes." *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D159–D162, Jan 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn772

[53] P. P. Chan and T. M. Lowe, „Gtrnadb: a database of transfer rna genes detected in genomic sequence." *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D93–D97, Jan 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn787

[54] T. Abe, T. Ikemura, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A. Muto, and H. Inokuchi, „trnadb-ce: trna gene database curated manually by experts." *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D163–D168, Jan 2009. [Online]. Available: http://dx.doi.org/10.1093/nar/gkn692

[55] T. Abe, T. Ikemura, J. Sugahara, A. Kanai, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A. Muto, and H. Inokuchi, „trnadb-ce 2011: trna gene database curated manually by experts." *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D210–D213, Jan 2011. [Online]. Available: http://dx.doi.org/10.1093/nar/gkq1007

[56] K. K. Kinouchi M, „trnafinder: A software system to find all trna genes in the dna sequence based on the cloverleaf secondary structure." *J. Comput. Aided Chem.*, vol. 7, p. 116–126, 2006.

[57] J. Sugahara, N. Yachie, Y. Sekine, A. Soma, M. Matsui, M. Tomita, and A. Kanai, „Splits: a new program for predicting split and intron-containing trna genes at the genome level." *In Silico Biol*, vol. 6, no. 5, pp. 411–418, 2006.

[58] J. Sugahara, K. Kikuta, K. Fujishima, N. Yachie, M. Tomita, and A. Kanai, „Comprehensive analysis of archaeal trna genes reveals rapid increase of trna introns in the order thermoproteales." *Mol Biol Evol*, vol. 25, no. 12, pp. 2709–2716, Dec 2008. [Online]. Available: http://dx.doi.org/10.1093/molbev/msn216

[59] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, „Information content of binding sites on nucleotide sequences." *J Mol Biol*, vol. 188, no. 3, pp. 415–431, Apr 1986.

[60] T. D. Schneider and R. M. Stephens, „Sequence logos: a new way to display consensus sequences." *Nucleic Acids Res*, vol. 18, no. 20, pp. 6097–6100, Oct 1990.

[61] J. Gorodkin, L. J. Heyer, S. Brunak, and G. D. Stormo, „Displaying the information contents of structural rna alignments: the structure logos." *Comput Appl Biosci*, vol. 13, no. 6, pp. 583–586, Dec 1997.

[62] C. E. Shannon, „A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[63] E. Freyhult, Y. Cui, O. Nilsson, and D. H. Ardell, „New computational methods reveal trna identity element divergence between proteobacteria and cyanobacteria." *Biochimie*, vol. 89, no. 10, pp. 1276–1288, Oct 2007. [Online]. Available: http://dx.doi.org/10.1016/j.biochi.2007.07.013

[64] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[65] G. P. Basharin, „On a statistical estimate for the entropy of a sequence of independent random variables." *Theory Probability Appl.*, vol. 4, no. 3, pp. 333–336, 1959.

[66] E. Jakó, P. Ittzés, A. Szenes, A. Kun, E. Szathmáry, and G. Pál, „In silico detection of trna sequence features characteristic to aminoacyl-trna synthetase class membership." *Nucleic Acids Res*, vol. 35, no. 16, pp. 5593–5609, 2007. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm598

[67] C. Marck and H. Grosjean, „trnomics: analysis of trna genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features." *RNA*, vol. 8, no. 10, pp. 1189–1232, Oct 2002.

[68] P. O'Donoghue and Z. Luthey-Schulten, „On the evolution of structure in aminoacyl-trna synthetases." *Microbiol Mol Biol Rev*, vol. 67, no. 4, pp. 550–573, Dec 2003.

[69] A. Ambrogelly, D. Korencic, and M. Ibba, „Functional annotation of class i lysyl-trna synthetase phylogeny indicates a limited role for gene transfer." *J Bacteriol*, vol. 184, no. 16, pp. 4594–4600, Aug 2002.

[70] J. D. Thompson, D. G. Higgins, and T. J. Gibson, „Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific

gap penalties and weight matrix choice." *Nucleic Acids Res*, vol. 22, no. 22, pp. 4673–4680, Nov 1994.

[71] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, „Multiple sequence alignment with the clustal series of programs." *Nucleic Acids Res*, vol. 31, no. 13, pp. 3497–3500, Jul 2003.

[72] D. M. Crothers, T. Seno, and G. Söll, „Is there a discriminator site in transfer rna?" *Proc Natl Acad Sci U S A*, vol. 69, no. 10, pp. 3063–3067, Oct 1972.

[73] W. H. McClain, K. Foss, R. A. Jenkins, and J. Schneider, „Rapid determination of nucleotides that define trna(gly) acceptor identity." *Proc Natl Acad Sci U S A*, vol. 88, no. 14, pp. 6147–6151, Jul 1991.

[74] T. Meinnel, Y. Mechulam, C. Lazennec, S. Blanquet, and G. Fayat, „Critical role of the acceptor stem of trnas(met) in their aminoacylation by escherichia coli methionyl-trna synthetase." *J Mol Biol*, vol. 229, no. 1, pp. 26–36, Jan 1993. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1993.1005

[75] O. Nureki, T. Niimi, T. Muramatsu, H. Kanno, T. Kohno, C. Florentz, R. Giegé, and S. Yokoyama, „Molecular recognition of the identity-determinant set of isoleucine transfer rna from escherichia coli." *J Mol Biol*, vol. 236, no. 3, pp. 710–724, Feb 1994. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1994.1184

[76] K. C. Rogers and D. Söll, „Discrimination among trnas intermediate in glutamate and glutamine acceptor identity." *Biochemistry*, vol. 32, no. 51, pp. 14 210–14 219, Dec 1993.

[77] M. Ibba, K. W. Hong, J. M. Sherman, S. Sever, and D. Söll, „Interactions between trna identity nucleotides and their recognition sites in glutaminyl-trna synthetase determine the cognate amino acid affinity of the enzyme." *Proc Natl Acad Sci U S A*, vol. 93, no. 14, pp. 6953–6958, Jul 1996.

[78] W. Freist, D. H. Gauss, M. Ibba, and D. Söll, „Glutaminyl-trna synthetase." *Biol Chem*, vol. 378, no. 10, pp. 1103–1117, Oct 1997.

[79] J. Normanly, T. Ollick, and J. Abelson, „Eight base changes are sufficient to convert a leucine-inserting trna into a serine-inserting trna." *Proc Natl Acad Sci U S A*, vol. 89, no. 12, pp. 5680–5684, Jun 1992.

[80] W. H. McClain and K. Foss, „Nucleotides that contribute to the identity of escherichia coli trna(phe).” *J Mol Biol*, vol. 202, no. 4, pp. 697–709, Aug 1988.

[81] Y. M. Hou, E. Westhof, and R. Giegé, „An unusual rna tertiary interaction has a role for the specific aminoacylation of a transfer rna.” *Proc Natl Acad Sci U S A*, vol. 90, no. 14, pp. 6776–6780, Jul 1993.

[82] R. S. Lipman and Y. M. Hou, „Aminoacylation of trna in the evolution of an aminoacyl-trna synthetase.” *Proc Natl Acad Sci U S A*, vol. 95, no. 23, pp. 13 495–13 500, Nov 1998.

[83] L. A. Sylvers, K. C. Rogers, M. Shimizu, E. Ohtsuka, and D. Söll, „A 2-thiouridine derivative in trnaglu is a positive determinant for aminoacylation by escherichia coli glutamyl-trna synthetase.” *Biochemistry*, vol. 32, no. 15, pp. 3836–3841, Apr 1993.

[84] S. Sekine, O. Nureki, K. Sakamoto, T. Niimi, M. Tateno, M. Go, T. Kohno, A. Brisson, J. Lapointe, and S. Yokoyama, „Major identity determinants in the "augmented d helix" of trna(glu) from escherichia coli.” *J Mol Biol*, vol. 256, no. 4, pp. 685–700, Mar 1996.

[85] S. Sekine, O. Nureki, M. Tateno, and S. Yokoyama, „The identity determinants required for the discrimination between trnaglu and trnaasp by glutamyl-trna synthetase from escherichia coli.” *Eur J Biochem*, vol. 261, no. 2, pp. 354–360, Apr 1999.

[86] M. J. Hohn, H.-S. Park, P. O'Donoghue, M. Schnitzbauer, and D. Söll, „Emergence of the universal genetic code imprinted in an rna record.” *Proc Natl Acad Sci U S A*, vol. 103, no. 48, pp. 18 095–18 100, Nov 2006. [Online]. Available: http://dx.doi.org/10.1073/pnas.0608762103

[87] N. Nameki, K. Tamura, H. Himeno, H. Asahara, T. Hasegawa, and M. Shimizu, „Escherichia coli trna(asp) recognition mechanism differing from that of the yeast system.” *Biochem Biophys Res Commun*, vol. 189, no. 2, pp. 856–862, Dec 1992.

[88] R. Giegé, C. Florentz, D. Kern, J. Gangloff, G. Eriani, and D. Moras, „Aspartate identity of transfer rnas.” *Biochimie*, vol. 78, no. 7, pp. 605–623, 1996.

[89] A. Fender, M. Sissler, C. Florentz, and R. Giegé, „Functional idiosyncrasies of trna isoacceptors in cognate and noncognate aminoacylation systems.” *Biochimie*, vol. 86, no. 1, pp. 21–29, Jan 2004. [Online]. Available: http://dx.doi.org/10.1016/j.biochi.2003.11.011

[90] H. Asahara, H. Himeno, K. Tamura, N. Nameki, T. Hasegawa, and M. Shimizu, „Discrimination among e. coli trnas with a long variable arm." *Nucleic Acids Symp Ser*, no. 29, pp. 207–208, 1993.

[91] N. Nameki, K. Tamura, H. Asahara, and T. Hasegawa, „Recognition of trna(gly) by three widely diverged glycyl-trna synthetases." *J Mol Biol*, vol. 268, no. 3, pp. 640–647, May 1997. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1997.0993

[92] N. Imura, G. B. Weiss, and R. W. Chambers, „Reconstitution of alanine acceptor activity from fragments of yeast trna-ala ii." *Nature*, vol. 222, no. 5199, pp. 1147–1148, Jun 1969.

[93] A. Soma, R. Kumagai, K. Nishikawa, and H. Himeno, „The anticodon loop is a major identity determinant of saccharomyces cerevisiae trna(leu)." *J Mol Biol*, vol. 263, no. 5, pp. 707–714, Nov 1996. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1996.0610

[94] J. Pütz, J. D. Puglisi, C. Florentz, and R. Giegé, „Identity elements for specific aminoacylation of yeast trna(asp) by cognate aspartyl-trna synthetase." *Science*, vol. 252, no. 5013, pp. 1696–1699, Jun 1991.

[95] M. Frugier, D. Söll, R. Giegé, and C. Florentz, „Identity switches between trnas aminoacylated by class i glutaminyl- and class ii aspartyl-trna synthetases." *Biochemistry*, vol. 33, no. 33, pp. 9912–9921, Aug 1994.

[96] J. R. Sampson, A. B. DiRenzo, L. S. Behlen, and O. C. Uhlenbeck, „Nucleotides in yeast trnaphe required for the specific recognition by its cognate synthetase." *Science*, vol. 243, no. 4896, pp. 1363–1366, Mar 1989.

[97] I. A. Nazarenko, E. T. Peterson, O. D. Zakharova, O. I. Lavrik, and O. C. Uhlenbeck, „Recognition nucleotides for human phenylalanyl-trna synthetase." *Nucleic Acids Res*, vol. 20, no. 3, pp. 475–478, Feb 1992.

[98] B. Mallick, J. Chakrabarti, S. Sahoo, Z. Ghosh, and S. Das, „Identity elements of archaeal trna." *DNA Res*, vol. 12, no. 4, pp. 235–246, 2005. [Online]. Available: http://dx.doi.org/10.1093/dnares/dsi008

[99] M. Naganuma, S.-i. Sekine, R. Fukunaga, and S. Yokoyama, „Unique protein architecture of alanyl-trna synthetase for aminoacylation, editing, and dimerization." *Proc Natl Acad Sci U S A*, vol. 106, no. 21, pp. 8489–8494, May 2009. [Online]. Available: http://dx.doi.org/10.1073/pnas.0901572106

[100] M. Sokabe, A. Okada, M. Yao, T. Nakashima, and I. Tanaka, „Molecular basis of alanine discrimination in editing site." *Proc Natl Acad Sci U S A*, vol. 102, no. 33, pp. 11 669–11 674, Aug 2005. [Online]. Available: http://dx.doi.org/10.1073/pnas.0502119102

[101] D. Tumbula-Hansen, L. Feng, H. Toogood, K. O. Stetter, and D. Söll, „Evolutionary divergence of the archaeal aspartyl-trna synthetases into discriminating and nondiscriminating forms." *J Biol Chem*, vol. 277, no. 40, pp. 37 184–37 190, Oct 2002. [Online]. Available: http://dx.doi.org/10.1074/jbc.M204767200

[102] E. Schmitt, L. Moulinier, S. Fujiwara, T. Imanaka, J. C. Thierry, and D. Moras, „Crystal structure of aspartyl-trna synthetase from pyrococcus kodakaraensis kod: archaeon specificity and catalytic mechanism of adenylate formation." *EMBO J*, vol. 17, no. 17, pp. 5227–5237, Sep 1998. [Online]. Available: http://dx.doi.org/10.1093/emboj/17.17.5227

[103] L. Feng, D. Tumbula-Hansen, H. Toogood, and D. Soll, „Expanding trna recognition of a trna synthetase by a single amino acid change." *Proc Natl Acad Sci U S A*, vol. 100, no. 10, pp. 5676–5681, May 2003. [Online]. Available: http://dx.doi.org/10.1073/pnas.0631525100

[104] K. Okamoto, A. Kuno, and T. Hasegawa, „Recognition sites of glycine trna for glycyl-trna synthetase from hyperthermophilic archaeon, aeropyrum pernix k1." *Nucleic Acids Symp Ser (Oxf)*, no. 49, pp. 299–300, 2005. [Online]. Available: http://dx.doi.org/10.1093/nass/49.1.299

[105] Y. Nagatoyo, J. Iwaki, S. Suzuki, A. Kuno, and T. Hasegawa, „Molecular recognition of histidine trna by histidyl-trna synthetase from hyperthermophilic archaeon, aeropyrum pernix k1." *Nucleic Acids Symp Ser (Oxf)*, no. 49, pp. 307–308, 2005. [Online]. Available: http://dx.doi.org/10.1093/nass/49.1.307

[106] W. Tsuchiya, M. Kimura, and T. Hasegawa, „Determination of phenylalanine trna recognition sites by phenylalanyl-trna synthetase from hyperthermophilic archaeon, aeropyrum pernix k1." *Nucleic Acids Symp Ser (Oxf)*, no. 51, pp. 367–368, 2007. [Online]. Available: http://dx.doi.org/10.1093/nass/nrm184

[107] J. Yokozawa, K. Okamoto, Y. Kawarabayasi, A. Kuno, and T. Hasegawa, „Molecular recognition of proline trna by prolyl-trna synthetase from hyperthermophilic archaeon, aeropyrum pernix k1." *Nucleic Acids Res Suppl*, no. 3, pp. 247–248, 2003.

[108] D. Korencic, C. Polycarpo, I. Weygand-Durasevic, and D. Söll, „Differential modes of transfer rnaser recognition in methanosarcina barkeri." *J Biol Chem*, vol. 279, no. 47, pp. 48780–48786, Nov 2004. [Online]. Available: http://dx.doi.org/10.1074/jbc.M408753200

[109] I. Gruic-Sovulj, J. Jaric, M. Dulic, M. Cindric, and I. Weygand-Durasevic, „Shuffling of discrete trnaser regions reveals differently utilized identity elements in yeast and methanogenic archaea." *J Mol Biol*, vol. 361, no. 1, pp. 128–139, Aug 2006. [Online]. Available: http://dx.doi.org/10.1016/j.jmb.2006.06.008

[110] Y. Nagaoka, J. Yokozawa, T. Umehara, J. Iwaki, K. Okamoto, Y. Kawarabayasi, Y. Koyama, Y. Sako, T. Wakagi, A. Kuno, and T. Hasegawa, „Molecular recognition of threonine trna by threonyl-trna synthetase from an extreme thermophilic archaeon, aeropyrum pernix k1." *Nucleic Acids Res Suppl*, no. 2, pp. 81–82, 2002.

[111] J. Yokozawa, Y. Nagaoka, T. Umehara, J. Iwaki, Y. Kawarabayasi, Y. Koyama, Y. Sako, T. Wakagi, A. Kuno, and T. Hasegawa, „Recognition of trna by aminoacyl-trna synthetase from hyperthermophilic archaea, aeropyrum pernix k1." *Nucleic Acids Res Suppl*, no. 1, pp. 117–118, 2001.

[112] H. Ishikura, Y. Nagaoka, J. Yokozawa, T. Umehara, A. Kuno, and T. Hasegawa, „Threonyl-trna synthetase of archaea: importance of the discriminator base in the aminoacylation of threonine trna." *Nucleic Acids Symp Ser*, no. 44, pp. 83–84, 2000.

[113] W. Tsuchiya, T. Umehara, A. Kuno, and T. Hasegawa, „Determination of tryptophan trna recognition sites for tryptophanyl-trna synthetase from hyperthermophilic archaeon, aeropyrum pernix k1." *Nucleic Acids Symp Ser (Oxf)*, no. 48, pp. 185–186, 2004. [Online]. Available: http://dx.doi.org/10.1093/nass/48.1.185

[114] J. Iwaki, H. Asahara, Y. Nagaoka, J. Yokozawa, T. Umehara, Y. Kawarabayasi, Y. Koyama, Y. Sako, A. Kuno, and T. Hasegawa, „Differences in tyrosine trna identity between escherichia coli and archaeon, aeropyrum pernix k1." *Nucleic Acids Res Suppl*, no. 2, pp. 225–226, 2002.

[115] J. Iwaki, R. Suzuki, Z. Fujimoto, M. Momma, A. Kuno, and T. Hasegawa, „Overexpression, purification and crystallization of tyrosyl-trna synthetase from the hyperthermophilic archaeon aeropyrum pernix k1." *Acta Crystallogr Sect F Struct Biol*

Cryst Commun, vol. 61, no. Pt 11, pp. 1003–1005, Nov 2005. [Online]. Available: http://dx.doi.org/10.1107/S1744309105033245

[116] T. Hasegawa, M. Miyano, H. Himeno, Y. Sano, K. Kimura, and M. Shimizu, „Identity determinants of e. coli threonine trna." *Biochem Biophys Res Commun*, vol. 184, no. 1, pp. 478–484, Apr 1992.

[117] N. Nameki, „Identity elements of trna(thr) towards saccharomyces cerevisiae threonyl-trna synthetase." *Nucleic Acids Res*, vol. 23, no. 15, pp. 2831–2836, Aug 1995.

[118] N. Nameki, H. Asahara, and T. Hasegawa, „Identity elements of thermus thermophilus trna(thr)." *FEBS Lett*, vol. 396, no. 2-3, pp. 201–207, Nov 1996.

[119] M. Pak, L. Pallanck, and L. H. Schulman, „Conversion of a methionine initiator trna into a tryptophan-inserting elongator trna in vivo." *Biochemistry*, vol. 31, no. 13, pp. 3303–3309, Apr 1992.

[120] M. Pak, I. M. Willis, and L. H. Schulman, „Analysis of acceptor stem base pairing on trna(trp) aminoacylation and function in vivo." *J Biol Chem*, vol. 269, no. 3, pp. 2277–2282, Jan 1994.

[121] M. J. Rogers, T. Adachi, H. Inokuchi, and D. Söll, „Switching trna(gln) identity from glutamine to tryptophan." *Proc Natl Acad Sci U S A*, vol. 89, no. 8, pp. 3463–3467, Apr 1992.

[122] Y. M. Hou and P. Schimmel, „A simple structural feature is a major determinant of the identity of a transfer rna." *Nature*, vol. 333, no. 6169, pp. 140–145, May 1988. [Online]. Available: http://dx.doi.org/10.1038/333140a0

[123] W. H. McClain and K. Foss, „Changing the identity of a trna by introducing a g-u wobble pair near the 3' acceptor end." *Science*, vol. 240, no. 4853, pp. 793–796, May 1988.

[124] L. Moulinier, S. Eiler, G. Eriani, J. Gangloff, J. C. Thierry, K. Gabriel, W. H. McClain, and D. Moras, „The structure of an asprs-trna(asp) complex reveals a trna-dependent control mechanism." *EMBO J*, vol. 20, no. 18, pp. 5290–5301, Sep 2001. [Online]. Available: http://dx.doi.org/10.1093/emboj/20.18.5290

[125] M. Ruff, S. Krishnaswamy, M. Boeglin, A. Poterszman, A. Mitschler, A. Podjarny, B. Rees, J. C. Thierry, and D. Moras, „Class ii aminoacyl transfer rna synthetases: crystal structure

of yeast aspartyl-trna synthetase complexed with trna(asp)." *Science*, vol. 252, no. 5013, pp. 1682–1689, Jun 1991.

[126] H. Liu and K. Musier-Forsyth, „Escherichia coli proline trna synthetase is sensitive to changes in the core region of trna(pro)." *Biochemistry*, vol. 33, no. 42, pp. 12708–12714, Oct 1994.

[127] R. Giegé, „Toward a more complete view of trna biology." *Nat Struct Mol Biol*, vol. 15, no. 10, pp. 1007–1014, Oct 2008. [Online]. Available: http://dx.doi.org/10.1038/nsmb.1498

Fajok	Adatbázisból letöltött szekvenciák száma (kiindulási adatok)	szekvenciák száma a ”kingdom” specifikus elemek szűrése után (első szűrési lépés)	szekvenciák száma az *E. coli* specifikus elemek szűrése után (második szűrési lépés)
Acetobacter_aceti	2	2	2
Acholeplasma_laidlawii	25	24	12
Acidithiobacillus_ferrooxidans	2	2	2
Acinetobacter_sp._ADP1	37	37	21
Aeromonas_hydrophila	7	6	5
Agrobacterium_tumefaciens	1	1	1
Agrobacterium_tumefaciens_str._C58	43	43	23
Aquifex_aeolicus_VF5	40	40	25
Azoarcus_sp._BH72	2	1	1
Azorhizobium_caulinodans	1	1	1
Azospirillum_lipoferum	3	3	1
Bacillus_anthracis_str._A2012	21	19	7
Bacillus_anthracis_str._Ames	55	54	30
Bacillus_anthracis_str._Sterne	60	59	31
Bacillus_cereus_ATCC_10987	60	60	36
Bacillus_cereus_ATCC_14579	53	53	30
Bacillus_circulans	1	1	1
Bacillus_halodurans_C-125	44	44	30
Bacillus_sp._PS3	5	3	2
Bacillus_subtilis	42	40	29
Bacillus_subtilis_subsp._subtilis_str._168	53	52	34
Bacillus_thuringiensis_serovar_konkukian_str._97-27	48	48	30
Bacteroides_thetaiotaomicron_VPI-5482	52	52	29
Bartonella_bacilliformis	1	1	1
Bartonella_elizabethae	2	2	2
Bartonella_henselae	1	1	1
Bartonella_henselae_str._Houston-1	40	39	24
Bartonella_quintana	2	2	2
Bartonella_quintana_str._Toulouse	38	37	22
Bdellovibrio_bacteriovorus_HD100	35	34	23
Bifidobacterium_longum_NCC2705	54	50	26
Bordetella_pertussis	1	1	1
Bordetella_sp.	1	0	0
Borrelia_burgdorferi	33	32	20
Borrelia_burgdorferi_B31	33	33	21
Borrelia_garinii_PBi	31	31	20
Bradyrhizobium_japonicum_USDA_110	48	47	29
Brucella_abortus	4	4	3
Brucella_melitensis	2	2	2
Brucella_melitensis_16M	46	46	29
Brucella_suis	2	2	1
Brucella_suis_1330	39	39	24
Buchnera_aphidicola_str._APS_(Acyrthosiphon_pisum)	31	31	25
Buchnera_aphidicola_str._Bp_(Baizongia_pistaciae)	32	32	23
Buchnera_aphidicola_str._Sg_(Schizaphis_graminum)	32	32	25
Burkholderia_cepacia	4	4	2
Burkholderia_gladioli	2	2	1
Burkholderia_mallei	2	2	1
Campylobacter_jejuni	6	6	5
Campylobacter_jejuni_subsp._jejuni_NCTC_11168	33	33	21
Candidatus_Blochmannia_floridanus	36	36	27
Candidatus_Protochlamydia_amoebophila_UWE25	34	34	21
Caulobacter_crescentus_CB15	43	42	22
Caulobacter_vibrioides	2	2	1
Chlamydia_muridarum_Nigg	37	37	21

Chlamydia_trachomatis	39	39	2
Chlamydia_trachomatis_D/UW-3/CX	36	36	21
Chlamydophila_pneumoniae_AR39	37	37	22
Chlamydophila_pneumoniae_CWL029	37	37	22
Chlamydophila_pneumoniae_J138	37	37	22
Chlorobium_tepidum_TLS	47	47	27
Clostridium_acetobutylicum	1	1	0
Clostridium_acetobutylicum_ATCC_824	42	42	26
Clostridium_perfringens	1	1	1
Clostridium_perfringens_str._13	40	40	26
Clostridium_tetani_E88	39	39	25
Corynebacterium_diphtheriae_NCTC_13129	44	44	21
Corynebacterium_efficiens_YS-314	49	49	23
Corynebacterium_glutamicum_ATCC_13032	49	49	24
Coxiella_burnetii	2	2	1
Coxiella_burnetii_RSA_493	42	41	30
Deinococcus_radiodurans_R1	45	44	26
Desulfomicrobium_baculatum	0	0	0
Desulfovibrio_vulgaris_subsp._vulgaris_str._Hildenborough	49	48	38
Enterococcus_faecalis_V583	48	47	26
Enterococcus_hirae	1	1	1
Escherichia_coli	44	44	40
Escherichia_coli_CFT073	51	48	38
Escherichia_coli_K12	46	45	39
Escherichia_coli_O157:H7	59	59	46
Escherichia_coli_O157:H7_EDL933	61	60	49
Fusobacterium_nucleatum_subsp._nucleatum_ATCC_25586	33	33	23
Geobacillus_stearothermophilus	1	1	0
Geobacter_sulfurreducens_PCA	43	43	33
Gluconacetobacter_europaeus	2	2	2
Gluconacetobacter_hansenii	2	2	2
Gluconacetobacter_liquefaciens	4	4	4
Gluconacetobacter_xylinus	2	2	2
Gluconobacter_oxydans	2	2	2
Haemophilus_ducreyi	1	1	1
Haemophilus_influenzae	48	48	34
Haemophilus_influenzae_Rd_KW20	39	39	29
Helicobacter_hepaticus_ATCC_51449	13	13	9
Helicobacter_pylori	35	35	22
Helicobacter_pylori_26695	35	34	22
Helicobacter_pylori_J99	35	34	22
Klebsiella_aerogenes	1	1	1
Lactobacillus_acidophilus	2	1	1
Lactobacillus_casei	2	2	2
Lactobacillus_curvatus	2	2	1
Lactobacillus_delbrueckii	1	1	0
Lactobacillus_delbrueckii_subsp._bulgaricus	10	10	7
Lactobacillus_helveticus	2	2	2
Lactobacillus_johnsonii_NCC_533	52	48	24
Lactobacillus_plantarum_WCFS1	52	49	26
Lactococcus_lactis	9	9	5
Leifsonia_xyli_subsp._xyli_str._CTCB07	45	42	20
Leptospira_interrogans_serovar_Lai_str._56601	36	36	26
Leuconostoc_lactis	1	1	1
Leuconostoc_mesenteroides	1	1	1
Listeria_innocua_Clip11262	49	49	33
Listeria_ivanovii	2	2	2
Listeria_monocytogenes	2	2	2
Listeria_monocytogenes_EGD-e	45	45	31
Listeria_monocytogenes_str._4b_F2365	44	44	31
Mesoplasma_florum_L1	26	26	12
Mesorhizobium_loti_MAFF303099	46	46	27

Moorella_thermoacetica	0	0	0
Mycobacterium_bovis_AF2122/97	44	44	22
Mycobacterium_leprae	3	3	1
Mycobacterium_leprae_TN	45	45	22
Mycobacterium_tuberculosis	2	2	1
Mycobacterium_tuberculosis_CDC1551	44	44	22
Mycobacterium_tuberculosis_H37Rv	45	45	22
Mycoplasma_capricolum	27	27	13
Mycoplasma_gallisepticum_R	31	31	16
Mycoplasma_genitalium	34	32	16
Mycoplasma_genitalium_G37	34	33	18
Mycoplasma_mycoides	14	14	6
Mycoplasma_pneumoniae	32	32	14
Mycoplasma_pneumoniae_M129	35	35	18
Mycoplasma_pulmonis_UAB_CTIP	27	27	13
Mycoplasma_sp.	1	1	1
Mycoplasma_sp._PG50	2	2	2
Neisseria_meningitidis_MC58	46	46	30
Neisseria_meningitidis_Z2491	37	37	24
Nitrosomonas_europaea_ATCC_19718	41	40	30
Nostoc_sp._PCC_7120	58	55	37
Oceanobacillus_iheyensis_HTE831	39	39	28
Ochrobactrum_anthropi	2	2	1
Onion_yellows_phytoplasma_OY-M	30	30	18
Pasteurella_multocida_subsp._multocida_str._Pm70	32	32	24
Photobacterium_leiognathi	2	2	2
Photobacterium_phosphoreum	2	2	2
Photorhabdus_luminescens_subsp._laumondii_TTO1	50	50	37
Phytoplasma_sp.	1	1	1
Plesiomonas_shigelloides	1	1	1
Porphyromonas_gingivalis_W83	46	45	24
Prevotella_ruminicola	2	2	2
Prochlorococcus_marinus	1	0	0
Propionibacterium_acnes_KPA171202	44	44	23
Pseudomonas_aeruginosa	7	7	4
Pseudomonas_aeruginosa_PAO1	42	41	31
Pseudomonas_fluorescens	2	2	1
Pseudomonas_mendocina	2	2	1
Pseudomonas_pseudoalcaligenes	2	2	1
Pseudomonas_putida	1	1	1
Pseudomonas_syringae_pv._tomato_str._DC3000	39	39	30
Ralstonia_pickettii	2	2	1
Ralstonia_solanacearum_GMI1000	47	47	32
Rhizobium_leguminosarum	1	1	1
Rhodopirellula_baltica_SH_1	72	63	36
Rhodopseudomonas_palustris_CGA009	46	45	29
Rhodothermus_marinus	2	2	0
Rickettsia_conorii_str._Malish_7	32	32	19
Rickettsia_prowazekii	3	3	3
Rickettsia_prowazekii_str._Madrid_E	33	33	19
Rickettsia_typhi_str._Wilmington	32	32	19
Ruminobacter_amylophilus	1	0	0
Salmonella_enterica_subsp._enterica_serovar_Typhi_str._CT18	48	48	42
Salmonella_enterica_subsp._enterica_serovar_Typhi_str._Ty2	48	48	41
Salmonella_enteritidis	2	2	2
Salmonella_typhimurium	5	5	5
Salmonella_typhimurium_LT2	50	50	43
Shewanella_oneidensis_MR-1	46	45	32
Shigella_flexneri_2a_str._301	58	58	43
Sinorhizobium_meliloti	2	2	2
Sinorhizobium_meliloti_1021	42	42	25
Spiroplasma_citri	3	3	2

Spiroplasma_melliferum		9	9	6
Staphylococcus_aureus		26	21	15
Staphylococcus_aureus_subsp._aureus_MRSA252		44	42	29
Staphylococcus_aureus_subsp._aureus_MW2		44	42	29
Staphylococcus_aureus_subsp._aureus_N315		42	40	29
Staphylococcus_epidermidis_ATCC_12228		47	44	29
Stenotrophomonas_maltophilia		2	2	1
Stigmatella_aurantiaca		4	4	3
Streptococcus_agalactiae_2603V/R		36	36	22
Streptococcus_agalactiae_NEM316		35	35	22
Streptococcus_mutans		1	1	1
Streptococcus_mutans_UA159		39	39	23
Streptococcus_pneumoniae		1	1	1
Streptococcus_pneumoniae_TIGR4		35	35	22
Streptococcus_pyogenes_M1_GAS		36	36	23
Streptococcus_pyogenes_MGAS315		34	34	20
Streptococcus_pyogenes_SSI-1		37	37	23
Streptococcus_salivarius		1	1	1
Streptomyces_ambofaciens		1	1	1
Streptomyces_avermitilis_MA-4680		57	55	26
Streptomyces_coelicolor_A3(2)		56	53	26
Streptomyces_griseus		2	2	2
Streptomyces_lividans		16	15	6
Streptomyces_rimosus		3	3	0
Streptomyces_venezuelae		1	1	1
Symbiobacterium_thermophilum_IAM_14863		67	67	43
Synechococcus_elongatus_PCC_6301		2	2	1
Synechococcus_sp.		3	3	2
Synechocystis_sp.		40	38	27
Synechocystis_sp._PCC_6803		40	38	27
Thermoanaerobacter_tengcongensis_MB4		52	52	32
Thermosynechococcus_elongatus_BP-1		41	39	27
Thermotoga_maritima		5	5	1
Thermotoga_maritima_MSB8		45	45	34
Thermus_thermophilus		5	5	4
Thermus_thermophilus_HB27		47	47	26
Treponema_denticola_ATCC_35405		41	40	23
Treponema_pallidum		44	44	24
Treponema_pallidum_subsp._pallidum_str._Nichols		44	44	24
Trichodesmium_sp.		2	2	1
Tropheryma_whipplei_str._Twist		45	44	23
Tropheryma_whipplei_TW08/27		46	45	23
Ureaplasma_parvum_serovar_3		28	27	14
Vibrio_cholerae_O1_biovar_El_Tor_str._N16961		51	50	34
Wolbachia_endosymbiont_of_Drosophila_melanogaster		34	34	23
Xanthomonas_axonopodis_pv._citri_str._306		47	47	33
Xanthomonas_campestris		1	1	0
Xanthomonas_campestris_pv._campestris_str._ATCC_33913		46	46	33
Xylella_fastidiosa_9a5c		69	68	45
Xylella_fastidiosa_Temecula1		45	45	27
Yersinia_pestis_biovar_Microtus_str._91001		46	46	36
Yersinia_pestis_CO92		45	45	36
Yersinia_pestis_KIM		48	48	38
Yersinia_pseudotuberculosis		1	1	1
Yersinia_pseudotuberculosis_IP_32953		45	45	35
	SUM	6243	6144	3901

Fajok	Adatbázisból letöltött szekvenciák száma (kiindulási adatok)	szekvenciák száma a "kingdom" specifikus elemek szűrése után (első szűrési lépés)	szekvenciák száma az élesztő specifikus elemek szűrése után (második szűrési lépés)
Arabidopsis_thaliana	215	198	154
Asterina_amurensis	1	0	0
Bombyx_mori	5	5	4
Bos_taurus	1	1	1
Brassica_napus	1	1	1
Caenorhabditis_elegans	218	184	143
Candida_albicans	4	3	3
Candida_cylindracea	1	0	0
Candida_glabrata_CBS_138	46	44	42
Candida_tropicalis	1	1	0
Clavispora_lusitaniae	1	0	0
Crithidia_fasciculata	2	2	2
Cyanidium_caldarium	3	2	1
Cyanophora_paradoxa	7	6	3
Dictyostelium_discoideum	20	19	18
Dromaius_novaehollandiae	1	0	0
Drosophila_melanogaster	123	122	99
Drosophila_simulans	2	2	2
Eimeria_tenella	1	1	0
Encephalitozoon_cuniculi_GB-M1	46	45	42
Euplotes_octocarinatus	1	1	1
Gallus_gallus	179	153	145
Glycine_max	3	2	2
Helianthus_annuus	1	1	1
Homo_sapiens	355	285	228
Leishmania_donovani	1	1	1
Leishmania_mexicana	1	1	1
Leishmania_tarentolae	10	10	9
Leptomonas_collosoma	3	3	3
Leptomonas_seymouri	2	2	2
Loligo_bleekeri	1	1	1
Lupinus_luteus	1	1	0
Mantoniella_squamata	1	1	1
Mus_musculus	13	11	11
Nephila_clavipes	4	2	2
Neurospora_crassa	3	3	3
Nicotiana_rustica	11	11	11
Oryctolagus_cuniculus	1	1	1
Oryza_sativa	1	1	0
Pan_troglodytes	305	245	228
Petunia_sp.	1	1	1
Phaseolus_vulgaris	5	5	5
Physarum_polycephalum	0	0	0
Phytophthora_parasitica	1	1	1
Pichia_guilliermondii	1	0	0
Plasmodium_falciparum	23	16	12
Plasmodium_falciparum_3D7	44	42	40
Podocoryne_carnea	5	5	4
Podospora_anserina	2	2	2
Pylaiella_littoralis	2	2	2
Rattus_norvegicus	25	25	25
Saccharomyces_cerevisiae	119	115	106
Salmo_salar	3	3	3
Schizosaccharomyces_pombe	93	88	64
Solanum_tuberosum	1	1	1
Sorghum_bicolor	1	1	0
Takifugu_rubripes	263	216	209
Tetrahymena_pyriformis	2	2	2
Tetrahymena_thermophila	2	2	2
Tinamus_tao	1	0	0
Toxoplasma_gondii	6	6	4
Triticum_aestivum	3	3	3
Trypanosoma_brucei	14	13	11
Xenopus_laevis	9	9	9
SUM	2222	1930	1672

Fajok	Adatbázisból letöltött szekvenciák száma (kiindulási adatok)	szekvenciák száma a "kingdom" specifikus elemek szűrése után (első szűrési lépés)
Aeropyrum pernix K1	36	36
Archaeoglobus fulgidus DSM 4304	23	18
Candidatus Methanoregula boonei 6A8	27	26
Cenarchaeum symbiosum	34	26
Haloarcula marismortui ATCC 43049	28	22
Halobacterium sp. NRC-1	25	21
Haloquadratum walsbyi	25	21
Hyperthermus butylicus DSM 5456	36	34
Ignicoccus hospitalis KIN4/I	24	24
Metallosphaera sedula DSM 5348	40	34
Methanobrevibacter smithii ATCC 35061	22	20
Methanocaldococcus jannaschii DSM 2661	25	22
Methanococcoides burtonii DSM 6242	25	24
Methanococcus aeolicus Nankai-3	29	25
Methanococcus maripaludis C5	31	26
Methanococcus maripaludis C7	31	26
Methanococcus maripaludis S2	31	26
Methanococcus vannielii SB	31	26
Methanocorpusculum labreanum Z	20	20
Methanoculleus marisnigri JR1	24	24
Methanopyrus kandleri AV19	28	23
Methanosaeta thermophila PT	26	22
Methanosarcina acetivorans C2A	31	29
Methanosarcina barkeri str. Fusaro	33	28
Methanosarcina mazei Go1	30	29
Methanosphaera stadtmanae DSM 3091	25	24
Methanospirillum hungatei JF-1	23	22
Methanothermobacter thermautotrophicus str. Delta H	22	20
Nanoarchaeum equitans Kin4-M	11	9
Natronomonas pharaonis DSM 2160	25	20
Picrophilus torridus DSM 9790	20	13
Pyrobaculum aerophilum str. IM2	33	31
Pyrobaculum arsenaticum DSM 13514	33	31
Pyrobaculum calidifontis JCM 11548	34	33
Pyrobaculum islandicum DSM 4184	34	32
Pyrococcus abyssi	39	36
Pyrococcus furiosus DSM 3638	35	31
Pyrococcus horikoshii OT3	36	33
Staphylothermus marinus F1	38	36
Sulfolobus acidocaldarius DSM 639	40	38
Sulfolobus solfataricus P2	39	38
Sulfolobus tokodaii str. 7	41	38
Thermococcus kodakaraensis KOD1	34	29
Thermofilum pendens Hrk 5	34	34
Thermoplasma acidophilum DSM 1728	14	8
Thermoplasma volcanium GSS1	14	8
uncultured methanogenic archaeon RC-I	29	24
Ismeretlen	184	164
SUM	1552	1384

2. melléklet

Bacteria				Eukaryota				Archaea	
	AEV				AEV				AEV
pozíció	második szűrési lépés nélkül	második szűrési lépéssel	NPD	pozíció	második szűrési lépés nélkül	második szűrési lépéssel	NPD	pozíció	nincs második szűrési lépés
0	8,05	6,2	1	0	1,9	2,75	0	0	5,15
1	12,55	11,15	5	1	10,45	10,85	2	1	5,2
2	12,5	10,95	9	2	10,3	9,8	1	2	10,05
3	7,05	10,05	7	3	7,9	7,85	2	3	11,4
4	2,65	6,9	5	4	8	6,9	0	4	10,25
5	1,85	5,05	1	5	6,8	5,15	0	5	10
6	1,85	5,2	0	6	2,75	1,9	0	6	7,6
7	5,15	7,3	0	7	5,7	5,65	0	7	8,05
8	0	0	1	8	0	0	0	8	4,55
9	8,05	9,5	0	9	5,7	6,4	0	9	7,55
10	9,05	7,6	2	10	7,85	8,5	1	10	4,9
11	6,35	6,85	2	11	4,35	2,75	0	11	10,8
12	6,7	10,55	1	12	9,75	7,35	0	12	11,7
13	9,95	11,9	2	13	10,3	10,05	0	13	8,7
14	5,05	3,55	1	14	0	0	0	14	0
15	9,15	7,95	1	15	6,05	6,15	0	15	0
16	5,35	7,45	0	16	8,15	7,75	0	16	9,45
17	6,9	8,2	0	17	9,5	9,3	0	17	8,05
17a	10,15	9,7	0	17a	4,7	6,3	0	17a	9,35
18	0	0	0	18	0	0	0	18	0
19	4,55	3,2	0	19	0	0	0	19	0
20	9,85	10,3	3	20	10,8	10,75	1	20	11,3
20a	9,4	10,65	0	20a	10,75	10,15	0	20a	10,8
20b	9,1	7,45	0	20b	7,85	9,15	0	20b	11,65
21	10,3	5,8	0	21	4,2	4,55	0	21	6,65
22	8,15	9,55	2	22	8,35	6,75	0	22	10
23	6,15	9,8	1	23	9,05	5,15	0	23	11,8
24	8,45	7,35	2	24	3,5	1,8	0	24	11,35
25	8,15	7,15	0	25	5,85	4,25	1	25	7,5
26	7,45	9	0	26	8,4	6,05	0	26	7,3
27	0,95	4,5	1	27	7,25	5,65	0	27	9,8
28	0,95	5,25	1	28	5,1	1,9	0	28	7,45
29	4,45	7,8	1	29	6	4,55	0	29	10,55
30	8	8,65	0	30	8,05	7,85	0	30	5,45
31	4,55	11,2	0	31	11,45	9,85	0	31	10,1
32	7,05	7,6	0	32	4,05	4,65	0	32	5,35
33	4,8	2,55	0	33	4,55	5	0	33	0
34	8,95	10	12	34	12,05	12,2	6	34	8,6
35	14,75	14,65	17	35	14,65	14,65	11	35	14,65
36	14,9	14,9	15	36	14,6	14,6	6	36	14,9
37	7,3	5,5	3	37	7,35	4,2	1	37	8,65
38	9,1	10,2	3	38	7,1	5	1	38	7,05
39	2,55	11,05	0	39	9,45	8,15	0	39	10,1
40	6,7	8,6	0	40	9,75	8,35	0	40	6,65
41	4,45	8,55	1	41	7,45	5,95	0	41	10,8
42	0,95	5,85	1	42	5,95	3,7	0	42	6,85
43	1,85	6,1	1	43	6,8	5,55	0	43	8,9
44	3,6	6,8	1	44	7,65	6,4	0	44	9,5
45	6,2	9	1	45	10,25	9	0	45	8,4
e11	4,85	4,3	0	e11	7,55	7,25	0	e11	-
e12	5	4,45	0	e12	6,5	7,15	0	e12	-
e13	4,9	3,35	0	e13	4,3	4,75	0	e13	-
e14	4,9	3,4	0	e14	3,6	3,65	0	e14	-
e15	3,4	1,8	0	e15	2,75	3,65	0	e15	-
e16	3,6	1,85	0	e16	1,85	1,85	0	e16	-
e17	1,85	1,85	0	e17	0,95	0,95	0	e17	-
e1	0,95	0,95	0	e1	0	0	0	e1	-
e2	8,5	7,4	0	e2	9,55	8,6	0	e2	-
e3	7,7	6,65	0	e3	8	7,8	0	e3	-
e4	5,7	4,3	0	e4	8,2	7,75	0	e4	-
e5	4,95	3,35	0	e5	5	5,55	0	e5	-
e27	1,85	1,85	0	e27	0,95	0,95	0	e27	-
e26	3,55	1,85	0	e26	1,85	1,85	0	e26	-
e25	3,5	1,8	0	e25	2,65	3,55	0	e25	-
e24	5	3,5	0	e24	3,6	3,45	0	e24	-
e23	4,85	3,35	0	e23	4,4	5,1	0	e23	-
e22	4,9	4,3	0	e22	6,5	6,95	0	e22	-
e21	4,85	4,3	0	e21	7,25	7	0	e21	-
46	8,6	10,75	1	46	9,3	8,9	0	46	10,9
47	11,55	11,3	1	47	11,25	10,85	0	47	5,6
48	7,85	6,85	1	48	8,15	8,25	0	48	0,95
49	5,15	8,3	0	49	7,2	5,15	0	49	8,65
50	0,95	3,65	0	50	7,35	6,45	0	50	7,7
51	7	8,65	0	51	9,25	8,35	0	51	8,1
52	7,55	6,75	0	52	7,4	5,95	0	52	10,9
53	0	0	0	53	0	0	0	53	0
54	0	0	0	54	6,85	8,15	0	54	0
55	0	0	0	55	0	0	0	55	0
56	4,2	3,2	0	56	0	0	0	56	0
57	0,95	0,95	0	57	0,95	0	0	57	0,95
58	0	0	0	58	0	0	0	58	0
59	1,85	5,1	1	59	6,65	5	0	59	8,55
60	7,5	7,85	1	60	7,2	7,9	0	60	0
61	0	0	0	61	0	0	0	61	0
62	7,05	6,45	0	62	6,4	6	0	62	9,4
63	5,2	8,4	0	63	8,6	7,8	0	63	7,45
64	0,95	4,55	0	64	9,05	8,6	0	64	6,7
65	5,15	7,6	0	65	6,6	5,9	0	65	8,95
66	4,55	7,3	0	66	5,35	5,35	0	66	7,3
67	1,85	5,2	0	67	4,45	4,25	0	67	8,85
68	3,5	5,15	1	68	4,85	1,85	0	68	10,35
69	0,95	7,6	5	69	6,8	4,3	0	69	10,5
70	6,55	11,3	7	70	9,2	8,15	2	70	12,1
71	10,2	10,4	9	71	10,45	8,15	1	71	10,2
72	9,8	10,2	7	72	9,8	9,55	2	72	5,95
73	11,4	12,25	18	73	9,2	12,05	9	73	10,4
átlag	5,59	6,45	1,63	**átlag**	6,31	5,79	0,49	**átlag**	7,34
szórás	3,49	3,53	3,51	**szórás**	3,47	3,41	1,69	**szórás**	3,95